

Chapter 1 – Probability and Statistics

1.1 - Our principal goal in this course will be to draw inferences about population parameters on the basis of sample statistics, to make decisions based on these inferences, and to quantify our confidence in these decisions. To do this, we use the mathematics of the theory of probability.

1.1.1 Definitions. The process of grouping, presenting, and quantifying the characteristics of sets of numbers is called *descriptive statistics*. Making predictions and decisions based on these data, and quantifying the *confidence* that we have in these predictions is called *statistical inference*. *Statistical analysis* can be defined as a set of methods for making decisions under uncertainty and for quantifying the confidence that we have in those decisions. The branch of mathematics that lays out the tools that we use in statistical analysis is called the *theory of probability*.

1.1.2 - Populations. A set of quantifiable data containing all values of interest in a particular analysis is called a *population*. Because the data must be quantifiable, if we are interested in looking at the wages of all workers at a construction site, the population is not the workers, but their wages. Dollars are quantifiable; people are not.

Sometimes it is easy to define a population. Wages of all workers on a job site would be easy to quantify because we know exactly who is there what their wages are. Other times, defining the population can be more difficult. Suppose the owner of the company is on the job site. His or her income might be disproportionately high. Moreover, it may not be possible to accurately determine the owner's "wages" because the owner might be salaried, receive bonuses, profits, etc.

We may have some control over the population. Suppose, for example, that we are pouring concrete for a bridge, and we are interested in implementing quality control by casting and breaking concrete cylinders made from concrete delivered to the job site. The population would be the set of all values of concrete cylinder strengths. However, because we design the quality control program, we can define the population by specifying the number of samples that must be taken, how they are to be tested, etc.

We need to be very careful to construct an appropriate population. In some cases, defining this population may involve considerable judgment on our part.

1.1.3 - Sampling. The parameters of a population are often determined from a subset of the population (or from another similar population) called a *sample*. Individual values from the sample are called *observations* or *sample points*. The numerical characteristics of the sample are often used to *estimate* the characteristics of the population. Any parameter of a sample used to estimate the characteristics of the population is called a *sample statistic*, or simply a *statistic*.

1.1.4 – Events. An event is a collection of one or more observations or sample points.

1.1.5 - Depicting Observations. Figure 1.1.5.1 shows a soil mass divided into a large number of hypothetical laboratory-sized blocks.

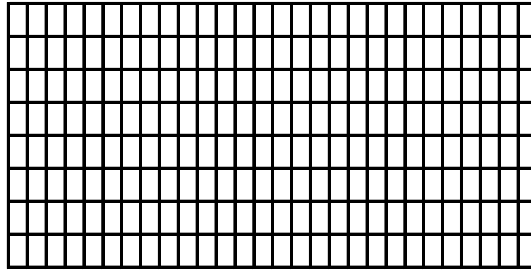


Figure 1.1.5.1 – Profile of ground showing one possible population of soil test specimens.

This array constitutes the population to be studied. It is not practical to test all of these blocks to determine their strength, therefore, we would select a convenient number samples, say 21, for laboratory testing. Table 1.1.5.1 shows a set of laboratory test results constituting the observations or sample points.

Test number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
cohesion, c	2180	1780	1620	1840	1960	2300	2040	2060	1850	1940	1840	1880	2045	1860	1945	2179	1760	1855	1920	1980	2070

TABLE 1.1.5.1 – Soil cohesion test values in #/ft²

Note that the soil cohesion values exhibit statistical variations. Some values are low, some high, and many are within an intermediate range around 2000 psf. This tabular view is not particularly helpful in visualizing the variability of the data.

To describe this variability, we will divide cohesion values into arbitrarily defined intervals of 100 psf, and tabulate the number of test values falling within each interval. Table 1.1.5.2 contains the intervals and the number of samples falling within each interval.

interval	below 1600	1600 to 1700	1700 to 1800	1800 to 1900	1900 to 2000	2000 to 2100	2100 to 2200	2200 to 2300	over 2300
number of tests	0	1	2	6	5	4	2	1	0

TABLE 1.1.5.2 - Data grouped into intervals of 100 #/ft² cohesive strength

Figure 1.1.5.2 presents a bar graph, called a histogram of observations, depicting the number of observations as a function of soil cohesion interval.

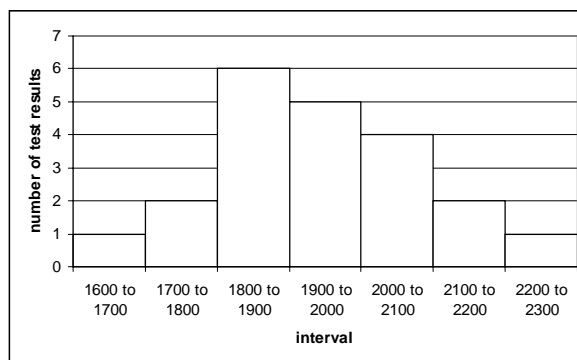


Figure 1.1.5.2 – Histogram of observations

This graphical relationship allows us to identify several characteristics of the data more clearly. First, we can see that the data clusters, or has a central tendency, around strengths of 1800 to 2000 psf. Second, we can see that the data is spread, or dispersed, over a range of several hundred psf. Third, we can see that the data is lopsided, being skewed to the right. In the next section, we will get more quantitative about these characteristics.

1.1.6 - Parameters. Numerical quantities developed to describe the characteristics of a population or of a sample from the population are called *parameters* of the population. Such parameters quantify important characteristics of the population. Consider the data for soil cohesion.

Central tendency. One important characteristic of a data set is its *central tendency*. Three measures of central tendency are commonly used. The *median* of a set of data is the value having as many observations greater than it as having observations less than it. For the above data, reference to Table 1.1.5.1 shows the median to be 1940 psf.

A second measure of central tendency is the *mode*. The mode is the value that occurs most frequently. Reference to Table 1.1.5.1 shows the mode to be 1840 psf, an observation that occurs 2 times.

A third measure of central tendency is the *mean*, which we will see later is numerically equal to the *average*. For this data set, the average is 1948 psf.

The median, the mode, and the mean all seem to give reasonable measures of central tendency. Both the median and the mean have the advantage that their values take into consideration all data points in the data set. The mode, however, may have been 1840 somewhat fortuitously in that there just happened to be two tests that gave that same result. Suppose that the 2179 test value had been 2180. How would this have affected the mode?

Dispersion. The spread of data about the central tendency is called the *dispersion*. One measure of dispersion is the *range* of the data, which is the number pair that represents the highest and lowest values. For the soil cohesion data set, the range is (1620, 2300).

A second measure of dispersion is the *absolute difference* taken as the highest value minus the lowest value. For this data set the absolute difference is $2300 - 1620 = 680$.

A third measure of dispersion that incorporates all data points and reflects the relationship of the dispersion to the mean is the *deviation*. If the value of any particular data point, i , is denoted by x_i , then the deviation of that data point from the mean is given by $(x_i - \text{mean})$. Because x_i can be above or below the mean, this is a signed number. A measure of the total deviation for n data points can, in principle, be determined by summing the individual deviations over all the points. This presents a problem because, given that the deviation is a signed number, negative values will cancel out positive values, artificially reducing the deviation. To circumvent this difficulty, the sign is obliterated by squaring the deviation. The result is the *variance*, defined as:

$$\text{variance} \equiv \frac{1}{n-1} \sum_{i=1}^n (x_i - \text{mean})^2$$

In order to get a measure of dispersion that has the same units as the data, the *standard deviation* is defined as:

$$\text{standard deviation} \equiv \sqrt{\text{variance}}$$

For this data set, the variance is 25,199. The standard deviation is 159.

Skewness. We also noted that the data was skewed to the right. A measure of skewness, given without supporting rationale, is:

$$\text{coefficient of skewness} \equiv \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \text{mean}}{\text{standard deviation}} \right)^3$$

The skewness for the soil cohesion data is 0.300, the positive value indicating that the observations are skewed to the right.

Later, we will look at the matter of parameters more comprehensively and develop a concept called the *method of moments* that will allow us to systematically define a set of parameters that, in the limit, can exactly characterize the data set. However, for now, the parameters defined above will suffice.

1.1.7 - Using Spreadsheets to Analyze Data. Small data sets can be analyzed with pencil and calculator; however, large data sets are better handled in a spreadsheet such as Microsoft Excel. What makes the spreadsheet so handy is that it has capabilities to plot the data and to determine common statistical parameters using built-in functions.

The observations are usually stored vertically in cells of a spreadsheet column. Column A often serves well. Because there are over 65,000 rows in a spreadsheet, a column can hold a lot of observations. As an example, enter the soil cohesion values in the spreadsheet range A1:A21.

Some useful parameters can be determined using formulas placed in spreadsheet cells. Some examples are:

1. Average, implemented as “=AVERAGE(A1:A21)” (don’t type the “=”) places the average (mean) value of the observations in the cell in which the formula is typed.
2. Mode, implemented as “=MODE(A1:A21)” places the mode of the observations in the cell in which the formula is typed.
3. Median, implemented as “=MEDIAN(A1:A21)”, places the median of the observations in the cell in which the formula is typed.
4. Standard deviation, implemented as “=STDEV(A1:A21) places the standard deviation of the observations in the cell in which the formula is typed.
5. Variance, implemented as “=VAR(A1:A21)” places the sample variance in the cell in which the formula is typed.
6. Skewness, implemented as “=SKEW(A1:A21)” places the skewness in the cell in which the formula is typed.

Excel also has some more comprehensive statistical analysis tools. Select the Tools menu item. There are several possible outcomes:

1. You may see a selection listed as Data Analysis... If you see this, you are in luck. We will discuss how to use it below.
2. You may not see Data Analysis..., but you will see Add-Ins... Select this and see if Analysis ToolPak – VBA is on the list of available add-ins. If so, turn its radio button on. The next time you Select Tools, you will find that Data Analysis along with several other new options will appear in the list.
3. If Analysis ToolPak – VBA is not available under the Add-Ins in (2) above, you need to use your original Excel (or Office) CD to do a custom install to load the Analysis ToolPak – VBA onto your hard drive. You can then go back to step (2) to make Data Analysis available on your Tools menu. You only have to go through steps (2) and/or (3) once. From then on, you will have the Analysis ToolPak components available on your Tools menu.

Assume that Data Analysis is available. Select Tools/Data Analysis and a Data Analysis window will pop up. You are interest in two items on this list: Descriptive Statistics and Histogram.

Descriptive Statistics is easy to use. Select it and click OK. A Descriptive Statistics window will pop up. Do the following:

1. In the Input Range, point to your data by entering \$A\$1:\$A\$21.
2. If it not already turned on, turn on the Columns radio button to tell Excel that your data is in columns.
3. Under Output Options, select New Spreadsheet Ply. This will create a new sheet in your workbook to which Excel will write the results. This will keep Excel from clobbering entries in other sheets in your workbook.
4. Click the check boxes for Summary statistics, Confidence Level for Mean, Kth largest, and Kth Smallest. (You might as well get everything Excel has to offer – there is no additional charge.)
5. Click OK to complete the process.
6. A new sheet will appear with your results on it.

Histogram is a little more complicated to use. We will use it in its simplest form first. Select Tools/Data Analysis/Histogram. Then do the following:

1. As before, specify the Input Range as \$A\$1:\$A\$21.
2. Under Output Options, turn on New Worksheet Ply if it is not already on.
3. Click the check boxes for Cumulative Percentage and for Chart Output.

4. Click OK.
5. A new sheet will appear with a chart on it.

The chart on the new sheet shows a histogram of observations for the data set. While the histogram looks pretty good, you will notice that the intervals that Excel chose for the bins is probably a strange number like 170. Normally, we like the bin end points to be nice round numbers. If Excel chose 170, we might like to change that to 100. To do this, we need to rerun Histogram and specify the bin break points ourselves rather than letting Excel specify them.

To accomplish this, we need to construct a *bin range*. The word range has a special meaning in Excel because it specifies a region on the spreadsheet. For our hypothetical data set, the range A1:A21 has special significance in that it is the range that contains the data. We now need to create a bin range: a range that contains the break points that we want for our bins. A good place would be a column other than column A; perhaps column B. We can start the bin range in cell B1 by typing a number in that cell that is (1) a multiple of 100, and (2) lower than the lowest observation in the data set. Specifically, type 1600 in cell B1. Successively type 1700, 1800, 1900, 2000, 2100, 2200, and 2300 in cells B2 through B8. Repeat the Histogram analysis as in steps 1 through 5 above except add the additional step of typing \$B\$1:\$B\$8 in the Bin Range input box on the Histogram window.

The information developed by Excel essentially duplicates the information presented in the first part of this chapter.

1.2 - Discrete and continuous sample spaces containing a finite or an infinite number of sample points.

A discrete sample space contains sample points (observations) that are countable using integers. Examples would be students in a classroom or grains of sand on the beach. A continuous sample space contains sample points that must be denoted by real (non-integer) numbers. An example would be the distance along a weld until the first flaw is encountered. All sample spaces can be classified as being either discrete or continuous. Sometimes the difference can be subtle. For example, the number of flaws in a given length of weld would constitute a discrete sample space; whereas the distance between consecutive flaws would be a continuous sample space. Why?

A discrete sample space can be finite or infinite. The number of students in a classroom would constitute a discrete finite sample space. The number of grains of sand on a beach would constitute a discrete infinite sample space. Sometimes the distinction between finite and infinite is difficult to ascertain. For example, do the grains of sand in a one gallon bucket constitute a finite or infinite discrete sample space? How about the number of grains of sand in a thimble? The answer to this question usually lies in examining the way in which we are using the sample space in our analysis. If we need to know or specify how many sample points there are in the sample space, then it will most likely be a finite one. If we must have 25 workers to do a job, then setting the problem as the probability that 25 workers will show up would involve an infinite sample space. That is, we wish to know if 25 out of a presumably infinite pool of workers will show up. On the other hand, if we set the problem in terms of the probability that no more than 5 workers will be absent from a staff of 30 workers, then the sample space would be finite. In both cases, the final result sought is the same: What is the probability that 25 workers will show up?

A continuous sample space will always be infinite. Assuming infinite resolution, the strength of a concrete cylinder can assume an infinite number of values (i.e. 1257.11793... psi). Moreover,

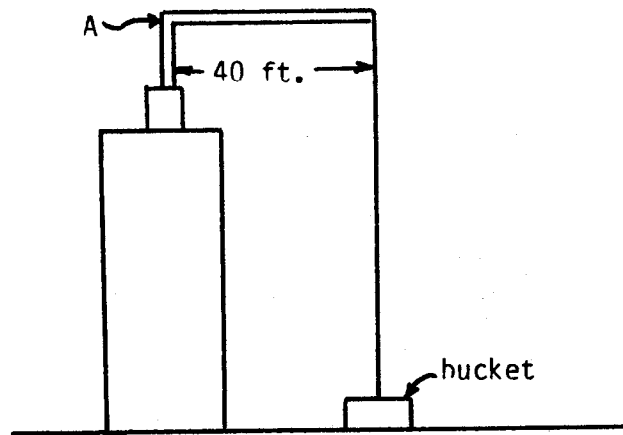
even if we limit the range of cylinders that are known to have strengths between 1000 and 1500 psi, there is still an infinite number of possible values within this range.

On the other hand, if the test machine has a digital readout displaying loads to the nearest pound, equipment limitations convert a continuous infinite sample space to a discrete finite one containing all integers from 1000 through 1500. Suppose that your test machine had an analog dial but you instructed your technician to record the test results to the nearest pound. What effect would this have on the sample space? Example Problem 1.1 illustrates the concepts presented in this section.

EXAMPLE PROBLEM 1.1

Discrete and continuous sample spaces of finite and infinite extent

Consider the crane shown below being used to lift construction materials at a building site. The crane lifts a bucket that is 6 feet long by 6 feet wide by 1 foot deep. The empty bucket weighs 500 pounds.



The moment at point A is given by $M=P(40)$, where P is the force in the cable. In this problem we will consider a sample space consisting of sample points giving values of M.

Initially, the elevator is used to raise bags of plaster. Each bag weighs 75 pounds and is 2 feet long by 1.5 feet wide by 1 foot high. Construct a logical sample space for M.

The lower bound for the sample space will be taken to be $500(40)=20\text{K ft}\cdot\#$. The upper bound will occur when the bucket is full of plaster bags. Assume that bags are laid 3 long ($3'\times 2'=6'$) by 4 wide ($4'\times 1.5'=6'$). Each such layer can be put in the bucket. Thus the bucket can hold 12 bags at most. The upper limit for $M=[12(75)+500][40]=56\text{K ft}\cdot\#$. Each bag contributes $75(4)=3\text{K ft}\cdot\#$ to the moment at point A. The sample space is a discrete finite space consisting of the sequence: 20, 23, 26 . . . , 56K ft-#.

Discuss factors that might distinguish this sample space from others in the possibility space.

- (1) We have assumed that the weight of the cable and boom are negligible.

EXAMPLE PROBLEM 1.1 (Cont.)

- (2) We have assumed the stacking arrangement. Other possibilities would include setting the bags on end or stacking the bags more than one layer deep.
- (3) We have assumed that no partial bags are loaded.

Can you list other assumptions?

The crane will also be used to lift bags of sand to be mixed with the plaster. The bags are the same size, but they weight 100 pounds each. The student should verify that the sample space for M , given that only sand bags are lifted, is discrete and finite and that it varies, therefore: 20, 24, 28, . . ., 68K ft-#.

After observing this operation, the foreman decides that it might be more efficient to: (1) only raise the bucket if it contains its full complement of 12 bags; and (2) allow the worker to fill the bucket with any combination of sand and/or plaster. Verify that the same space for M is discrete and infinite and that it varies therefore: 57.0, 58.0, 59.0, 60.0, 61.0, 62.0, 63.0, 64.0, 65.0, 66.0, 67.0, 68.0K ft-#.

The project superintendent walks by one day and suggests that the worker on the ground should open the bags, dump them into the bucket, and dry mix them. The mix proportions are one bag of plaster to one bag of sand. The bucket is still to contain 12 bags. Confirm that the sample space now consists of one sample point, $M = 62K$ ft-#.

The project union steward observes that the worker's work load is too heavy. He suggests that the bucket need not be full but that any combination of whole bags be allowed. The foreman agrees with this but reminds the worker that the one-to-one mix proportion must be maintained. The project statistical analyst, lamenting that no one seems to be concerned about his work load then defines the new sample space to be 27.0, 340.0, 41.0, 48.0, 55.0, 62.0K ft-#.

An efficiency expert than enters the picture and suggests that the worker also add the water to the mix. The water contributes no additional volume to the mix. Each bag of plaster requires 25 pounds of water. At this point, the plasterers irately proclaim that they wish that the people on the ground would get their act together. The plasterers say they never know that to expect in the bucket and, besides, the worker on the ground (not a certified plasterer) might get the mix too wet. Moreover, the worker on the ground would always be in the middle of mixing when they need plaster, thus holding up their work. The plasterers suggested the following procedure: The worker on the ground should put in 6 bags of plaster and 6 bags of sand. He should then add water—a little at a time—and mix the plaster. When the plasterers need more plaster, they will call for the mix as is and will complete the addition of water until the consistency is just right. Confirm that the sample space now becomes a continuous one (and therefore infinite), ranging from $62.0K \leq M \leq 68.0K$ ft-#.

1.3 - Depicting sample spaces using Venn diagrams.

The concepts presented in the preceding sections can be mathematically treated using set theory. It should be pointed out in advance that the focus of this book will not be toward solving problems using mathematically abstract set theory approaches, but rather we will use these principles to develop graphic approaches that are more appropriate to solving engineering problems.

Set theory uses the Venn diagram to depict sample spaces, events, and sample points. The Venn diagram is drawn in such a way that interrelationships among events can be meaningfully shown. As seen in Fig. 1.3.1, the sample space, (S), is denoted by a rectangle. Events such as E_1 and E_2 occupy regions within the sample space. Interrelationships among events are described by allowing the regions representing events to overlap when necessary.

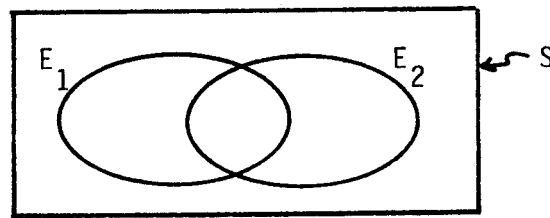


Figure 1.3.1 Venn diagram describing the sample space, S , and two events, E_1 and E_2

1.4 - Special events: impossible events, certain events, complementary events, mutually exclusive events, intersecting events, collectively exhaustive events, and included events.

In Section 1.1 we defined sample points, events, and sample spaces. In terms of set theory, a sample point is a subset of an event which is in turn a subset of the sample space. By examining the relationships among these entities we see the practicality of defining several additional terms. Example Problem 1.2 illustrates these new terms. An impossible event (denoted by ϕ) is an event that contains no sample points. By contrast, the certain event (denoted by S) is one that contains the entire sample space. Note that this definition is at odds with our everyday use of the term certain. Normally, we might expect this to mean an event that will definitely happen. In fact, the probability theory definition implies that whatever happens, if anything, must be contained within the sample space.

As shown in Fig. 1.4.1, if there is an event, E , in a sample space, the complementary event, \bar{E} , is the event that contains all sample points in the sample space not included in event E . If we undertake an action to create \bar{E} , we are said to be creating the complement of E .

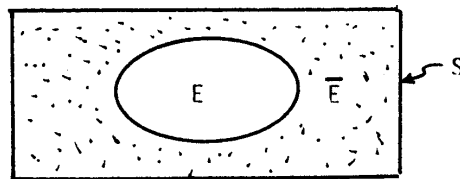


Figure 1.4.1 – An event E and its complement \bar{E}

Mutually exclusive events are events that do not have any sample points in common, as shown in Fig. 1.4.2.

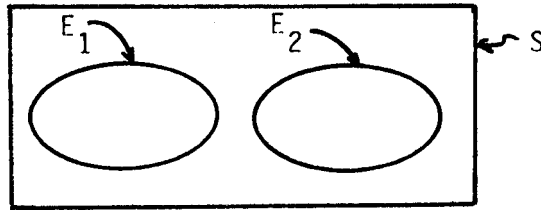


Figure 1.4.2 – Events E_1 and E_2 and mutually exclusive because they share no sample points.

Intersecting events are events that share one or more sample points, as shown in Fig. 1.4.3. The region that contains the shared sample points (the football) is called the intersection of E_1 and E_2 and is written as $E_1 \cap E_2$, or simply as $E_1 E_2$. Sample points that belong to the intersection are said to be in E_1 and E_2 .

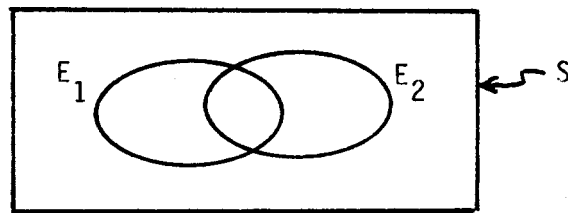


Figure 1.4.3 Events E_1 and E_2 intersect because they share sample points.

The union of events is created from the region that contains sample points that are in either E_1 or E_2 or both as shown by the owl eyes in Fig. 1.4.4. The union is written $E_1 \cup E_2$, and sample points that belong to the union are said to be in E_1 and/or E_2 .

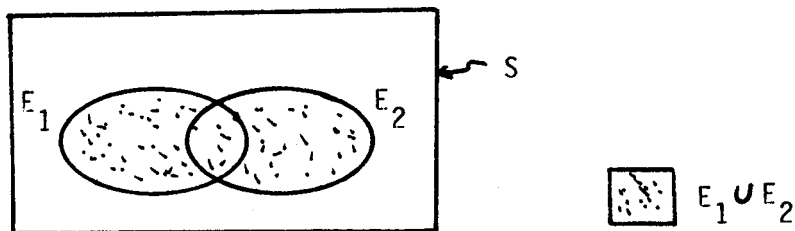


Figure 1.4.4 The shaded area is the union of E_1 and E_2 .

Collectively exhaustive events are events that, when taken together, contain all sample points in the sample space as shown in Fig. 1.4.5.

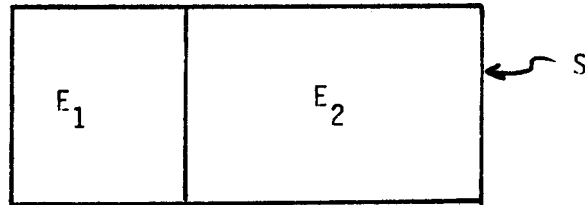


Figure 1.4.5 – E_1 and E_2 are collectively exhaustive because, taken together, they include all sample points in the sample space.

If all of the sample points in events E_2 are also contained in event E_1 , then E_2 is said to be an included event within E_1 . Figure 1.4.6 illustrates this.

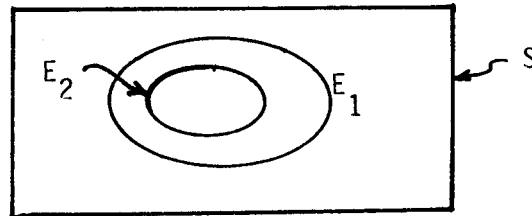


Figure 1.4.6 Event E_2 is included in event E_1 .

EXAMPLE PROBLEM 1.2

Events

A contractor has a fleet of three dump trucks available for use on a project. We would like to examine the probability of one or more of these vehicles being operational (O) or non-operational (\bar{O}) on a given day. In this example problem we will establish the same space and define sample points and events within the sample space. Calculation of probabilities will be reserved for later examples.

A sample point will be defined as a set of three ordered letters designating the operational state of all three trucks. Thus, sample point $O\bar{O}O$ denotes that truck #1 is operational, truck #2 is non-operational, and truck #3 is operational.

The sample space consists of all possible combinations of operational and non-operational states of vehicles. This sample space and the sample points of which it is composed are shown below.

EXAMPLE PROBLEM 1.2 (Cont.)

The sample space defines the events that we will recognize as possible in our analysis. However, the sample space may differ from the possibility space. For example, the contractor owns a fourth truck, but it is away at the factory being overhauled. We have excluded it from our analysis. I also happen to know that truck #3 runs reliably only in the first three gears. High gear often will not operate. Our analysis does not include this aspect. As far as we are considered, if it operates in three out of four gears, it operates. Thus, as pointed out in the text, the difference between the possibility space and the sample space serves to define what we have chosen to exclude from the analysis. Could you construct a sample space that also incorporates the unreliability of high gear in truck #3?

We will now define some events in the sample space:

- (1) Event E_1 is the event that trucks #1 and #2 operate but truck #3 does not. Note that this event contains only one sample point.
- (2) Event E_2 is the event that two or more trucks are non-operational. Can you define the event that exactly two trucks are non-operational? In this course the phrase “two trucks are non-operational” implies that exactly two trucks are non-operational. Can you define the event that more than two trucks are non-operational? Note that the three events—two trucks non-operational, two or more trucks non-operational, and more than 2 trucks non-operational—are separate and distinct events, even though the English phrases sound similar. Students should read problem wording carefully. There will be no attempt on my part to use trick wording in problems, however, careless reading can easily get you off on the wrong track on a problem. If you are not sure about your interpretation of wording, include a clarification in your solution. For example, you may wish to state that in solving this problem you assumed that the phrase “two trucks” meant “exactly two trucks.”
- (3) Event E_3 is the that truck #1 is operational. Note that the way in which we arranged the sample space makes it awkward to delineate E_3 . Can you arrange the sample space so that E_3 is more easily delineated? Note also that even though event E_3 only speaks directly about the status of truck number 1, the operational status of trucks #2 and #3 implicitly enters the picture.
- (4) The event E_4 that truck #1 is operational, truck #2 is non-operational, and the status of truck #3 is unknown is an impossible event in the sense defined in the text, even though the event can happen. Why? Is the event truck #1 is operational, truck #2 is non-operational and truck #3 is either operational or non-operational an event that can be defined in the sample space? Delineate it. How about the event that truck #2 is operational, truck #3 is non-operational, and truck #1 is both operational and non-operational?
- (5) The certain event is the entire sample space. In this example problem it contains eight sample points. Delineate it.
- (6) Event \bar{E}_3 is the complement of event E_3 . They are also mutually exclusive events. It is possible to define many mutually exclusive events for this example. Think up some of your own. Also think of some events that are not mutually exclusive.

EXAMPLE PROBLEM 1.2 (Cont.)

- (7) If event E_5 is that truck #1 is operational and truck #2 is non-operational, and if event E_6 is that truck #1 is non-operational and truck #2 is operational, then events E_5 and E_6 are mutually exclusive.
- (8) If event E_7 is that more than two trucks are operational, if event E_8 is that less than two trucks are operational, and if event E_9 is that exactly two trucks are operational, then E_7 , E_8 , and E_9 are collectively exhaustive. They also happen to be mutually exclusive.
- (9) If event E_{10} is that truck #1 is operational and E_{11} is that truck #2 is operational, then E_{10} and E_{11} are intersecting events. The common sample points are OOO and $OO\bar{O}$.
- (10) Event E_7 is included in event E_{10} . Identify other examples of included events in events E_1 through E_{11} .

The solution is shown in the following diagram:

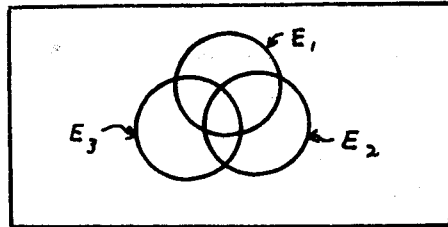
Truck 1	Truck 2	Truck 3	E_1	E_2	E_3	\bar{E}_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}
O	O	O			✓					✓			✓	✓
\bar{O}	O	O				✓			✓			✓		✓
O	\bar{O}	O			✓			✓				✓	✓	
O	O	\bar{O}	✓		✓							✓	✓	✓
\bar{O}	\bar{O}	O		✓		✓					✓			
O	\bar{O}	\bar{O}		✓	✓			✓			✓		✓	
\bar{O}	O	\bar{O}		✓		✓			✓		✓			✓
\bar{O}	\bar{O}	\bar{O}		✓		✓					✓			

EXAMPLE PROBLEM 1.3

Interrelating events using Venn Diagrams

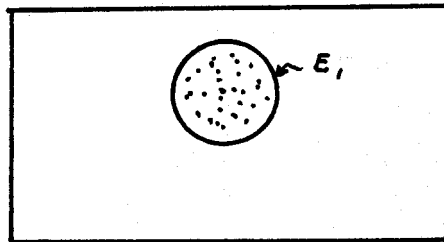
Consider three events E_1 , E_2 , and E_3 that exhibit the most general relationships possible among the events. Construct a Venn diagram.

The exercise is not as open ended as it might at first appear. There will be many occasions where we will need to construct a Venn diagram that allows for all possible interrelationships among events. As the analysis proceeds, it may then be possible to simplify the Venn diagram based on additional information. The figure below shows a highly versatile Venn diagram.

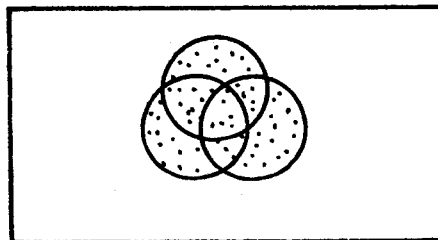



The student should identify the regions defining the following events:

- (1) Events E_1 , E_2 , and E_3 each are defined by a circle.



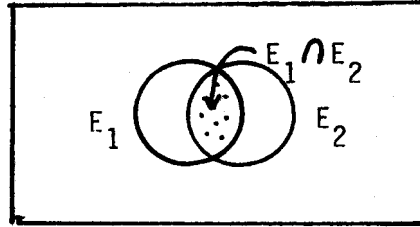
- (2) Event $E \equiv E_1 \cup E_2 \cup E_3$ defined by the three-leaf clover.



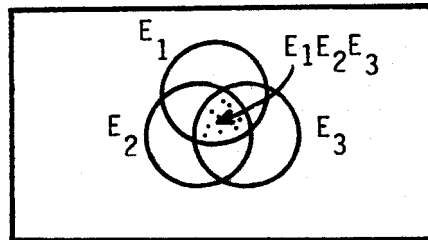
 $E_1 \cup E_2 \cup E_3$

EXAMPLE PROBLEM 1.3 (Cont.)

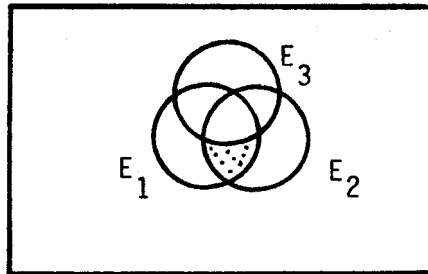
- (3) Events $E_1 \cap E_2$, $E_2 \cap E_3$, and $E_3 \cap E_1$, each defined by a football.



- (4) Event $E \equiv E_1 \cap E_2 \cap E_3 \equiv E_1 E_2 E_3$, defined by a Wankel engine.



- (5) Confirm that the event $E_1 \cap E_2 \cap \bar{E}_3$ is given by the highway patrol arm patch.



1.5 – Working word problems.

As with most disciplines, working word problems in probability can be tough. The basis for the difficulty is that the discipline makes a very limited vocabulary available to the analyst; yet people (clients) tend to use a more varied and less exact vocabulary to describe their problems. The job of the engineer is to translate the client's statements into the more limited vocabulary of the discipline.

So far, we only have four words in our vocabulary:

- (1) Names of events, like E_1 or $E_1 \cap E_2$ or the soil cohesion value 1040 psf.
- (2) The word not which denotes the complement of an event.
- (3) The word and which denotes the intersection of events.
- (4) The word and/or which denotes the union of events.

A simple example. A client might say: Pumps 1 and 2 operate, but pump 3 doesn't. This is a great English sentence, but a bad probability statement. What's wrong with it? First, "Pumps 1 and 2 operate" is vague. A clearer statement would be "pump 1 operates and pump 2 operates". Second, "but pump 3 doesn't" is full of vague and disallowed words. A clearer statement would be "pump 3 does not operate". The client's statement can be rewritten using only allowable vocabulary words as "pump 1 operates and pump 2 operates and pump 3 does not operate."

Now let's define some events: P_1 is pump 1 operates; P_2 is pump 2 operates, \bar{P}_3 is pump 3 does not operate. And is the intersection operator. We can now transform our restatement of the client's concerns into a theory of probability statement:

$$P_1 \cap P_2 \cap \bar{P}_3$$

Draw a Venn diagram using three circles to describe possible operating states for the pumps. Shade in the event described above.

The client also said that on another occasion, "pump 3 is the only pump that does not operate". Translate this into a sentence that only uses allowable words. Show the event on your Venn diagram.

Now try "one or more pumps operates". Show it on the Venn diagram.

Make up some more possible situations. Then translate them into allowable words and show them on the Venn diagram.

1.6 – Extending the concept of Venn diagrams to depict sample spaces in analytically convenient ways.

In the strict mathematical sense, a Venn diagram is limited to the type of figure shown in the section 1.4. However, we will extend the concept to include any pictorial representation showing a sample space in a way that aids in examining events and in quantifying the probability of their occurrence.

The first new type of diagram is the axial sample space. In its simplest form, it consists of a single axis. A continuous implementation, as shown in Fig. 1.6.1, can be used depict events such as the distance along a weld to the occurrence of a flaw.

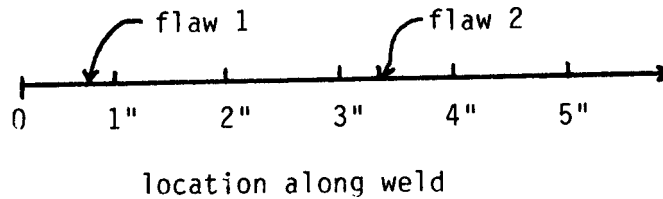


Figure 1.6.1 – Uniaxial, continuous sample space.

A discrete implementation could locate the position of a person in a ticket line.

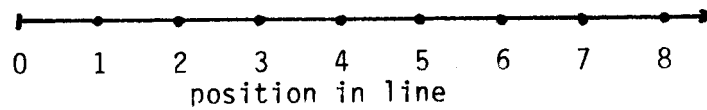


Figure 1.6.2 – Uniaxial, finite, discrete sample space.

A second type of diagram is the tabular sample space. The diagram used in Example Problem 1.2 is a form of tabular sample space with the table consisting of a one column table with eight sample point, each sample point representing one operating state for the three trucks. Fig. 1.6.3 shows a two-dimensional tabular sample space.

land use state	urban	farm
Ohio	100	80
Kentucky	90	120
Indiana	80	150

Figure 1.6.3 – A tabular sample space having two columns and three rows.