

Notes for MSIT 431 for Fall 2011

Statistical Methods¹ by A. H. Haddad

1.0 Introduction to Data Analysis

1.1 Introduction

Why do we need probability and statistics in networks, computers, and telecommunications? In most problems of information technology and telecommunications, we deal with large amounts of data and with many uncertainties that affect the performance and the cost of our systems. Probability is aimed at letting us design systems and processes that will perform well despite the presence of these uncertainties. In general with telecom systems we have large amounts of data dealing with the number of users in the systems at any one time, in the number of failures of branches or subsystems, in a variety of congestion conditions that may affect the performance as viewed by the customer. Many of these systems need to be designed by using as performance criteria some average property, rather than worst case design as is usually done with many other engineering systems. It means that such systems cannot usually cope with extreme conditions. For example when many customers try to call at the same time due to some natural disaster or a special event, usually most will not be able to make a connection. Similarly, if many users during rush hour in a large metropolitan area try to call while driving, it is very likely that some calls will be dropped off. However, how do we manage to get acceptable behavior, despite the fact that many of the variables that may cause changes in behavior are not precisely known? In such systems one cannot expect certainties. Almost nothing can be guaranteed to be available 100% of the time.

In addition, the need for statistical methods is also critical for the business and financial aspects of IT. For example, all usage is based on some average properties that may affect pricing. Using the incorrect approach may lead to one of the following extreme cases: either spending more than the business needs to get a desired performance or ending up with less than the business requires to meet customers demand resulting in many unhappy customers. Here too, when dealing with average properties, one may be disappointed when such averages do not materialize in reality. One always hears averages being bandied about, such as percentage of delinquent customers, percentages of credit card fraudulent transactions, or percentage of foreclosed mortgages. How does this average affect an individual business? Does size matter when using such averages? We shall see that averages are meaningful only in certain situations. One is used to averages when dealing with weather conditions, such as average rain fall or average high or low temperatures. In many cases weatherpersons may refer to the average as “normal” which is not a very meaningful term. It is important to know when the average is indeed a meaningful number that provides an indication of acceptable conditions. In the weather information case the weatherperson also provides the maximum and the minimum of such temperatures besides the averages.

¹Copyright 2011 A. H. Haddad

Examples of data types that may be available and which need to be analyzed in order to have a better system design may include:

- a. Records of calls or packets at a telecom switch
- b. Length of messages or service times of calls
- c. Records of failures of various branches and nodes in the network
- d. Records of error rates in transmitting blocks of data
- e. Records of number of packets dropped in a packet transmission system
- f. Lifetime of various branches or components of a system
- g. Amount of latency in time-critical messages
- h. Number of expected users at any particular period
- i. Records of dropped calls during handoff
- j. Number of delinquent credit-card customers
- k. Amounts owed by such delinquent customers
- l. Aggregate financial records of transactions
- m. Number of transactions involving certain products

Now if we collect such data we have to be able to analyze them in order to extract useful information, that helps determine system performance or system design. The analysis may help determine parameters of the systems, the need for certain capacity, the amount to charge customers for using the system, as well as help make decisions with reasonable amount of certainty of the outcome. The purpose of this analysis is to be able to design a system that works within given constraints, which are both technical and financial. Listed below are examples of such system design issues.

- a. How many lines should we have to serve two communications nodes?
- b. How many customers should a wireless base station be able to handle?
- c. What is the acceptable rate of dropping wireless calls during handoff?
- d. How to reroute packets when there are failures or congestions in some branches?
- e. How many outside lines a company needs to install to serve its employees?
- f. What is the maximum latency allowed and with what probability?
- g. When to self-insure against losses and when to use external insurance?
- h. What is the balance needed to ensure against delinquent loans?
- i. What other products to offer customers who bought certain products?
- j. How many servers are needed to obtain a desired probability of busy periods?

1.1.1 How do we use the data?

One may ask if data alone is used for analysis, or do we need additional information to help in the analysis. In order to make a reasonable attempt at answering some of these questions, we need to have systematic models that can be used together with the data. If we are only given data alone, we tend to use statistics, which does not assume any underlying models, but just derive a variety of means to represent and analyze the data. Since we are dealing with man-made systems (albeit very complex ones) in the information and telecommunications field, we can actually

provide mathematical models that together with the large amounts of data that is available may yield an excellent tool for the analysis and design of such systems. Such analyses are based on statistical models, and we use the data to verify the various assumptions and identify the various parameters in such models. This type of approach is called *inferential statistics*. For example, we design a system based on a model to have a given error rate. We then need to analyze the data provided by such system to verify that the desired error rate is indeed achieved.

In the social sciences and even some biomedical sciences and other business models one has to be satisfied with statistics that its primary purpose is the analysis of data and the determination of some relationships. That is the subject of pure statistics, where we collect data and analyze it to make critical choices or decisions. In engineering, we can actually supply good basic models based on physical laws or logical assumptions, so the analysis of the data is performed with these models in mind. In many cases we have to analyze data of both types. For example, collecting usage data and other financial aspects of the operation belongs to the “pure statistics” model, while the system performance in the presence of such data may belong to the “inferential statistics” model. Even without a given model, if we have very large amounts of data we can make serious assumption on what models fit such data best.

We start by first discussing the approaches used to analyze data and how such data is to be represented. Then we discuss mathematical models based on probability to construct reasonable models under some simple assumptions. Finally, we bring the two together to see how the models can be identified from data sets and how decisions are made based on the data.

1.2 Representation of Data

The first issue we discuss in this course is how to analyze data that we collect. What do we do with the huge amount of data available? A variety of graphical approaches are used to display and represent the data more compactly. We mention a few here as examples:

- a. Plot the data versus time. This will tell us if time is a key parameter.
- b. Plot the data as dots on a graph
- c. Stem plots (useful only for small amounts of data)
- d. Divide the data into ranges and plot the number in each range
- e. Box plots
- f. Bar graphs or stacked bar graphs
- g. Plots different types of data against each other to observe relationships
- h. Show the data in a two-way or three-way tables

The way we represent the data depends on its type. In general, we have two types of data: *categorical* and *numeric*. Other terminology may also be used for such distinction: *discrete* and *continuous*. Categorical (or discrete) data means that the data falls into distinct categories, for example, in describing the education level of a population, or the failure or non-failure of a system, or the quality of a product, or the letter grade of a student. Numeric or continuous data can take a range of numeric values, such as the signal level in a wireless system, or the income level in a population, or the numeric grade in an exam. Usually, it is possible by properly

partitioning the range of continuous data sets to convert such data into discrete. This can be done only if we have a large amount of data and if we do not wish to distinguish among the data in a given range. Both textbooks describe many ways of graphically representing data as listed above. We shall limit our coverage to just two graphical representations of data:

- a. **Histograms,**
- b. **Box plots.**

While we shall discuss both of these in some detail, in almost all engineering and technical problems, we usually prefer to use histograms and not box plots.

1.2.1. Histograms

What is a histogram and how do we construct a histogram of the data available?

If we divide the range of the data into a finite number of ranges and count how many of the points fall in each range, the resulting picture is called a histogram. There are several ways of plotting a histogram. We can use the actual number of data points in each range, so that the values on the vertical axis will depend on the total number of data points. The values will change if we also change the number or width of ranges. However, a more informative value is the relative frequency of the occurrence of data points in each range. The relative frequency approach is the more informative one, which is obtained by dividing the number in each range, by the total number of values in the sample. Another problem occurs if we try to use different sizes for the ranges. If we increase number of the ranges but make each smaller in size, the vertical size of the graph will get smaller as well. As an example, refer to Figure 1.1, where the histograms of 500 data points between 0 and 10 are shown using relative frequency with 10 and 5 ranges.

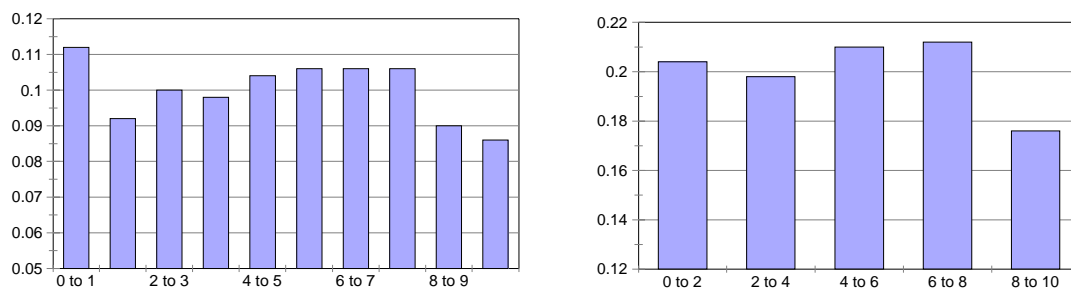


Figure 1.1 Histogram of 500 data points, using five and ten ranges of values and displaying relative frequency. To obtain a consistent histogram we divide the values by 2 in the figure on the right, so that the area each bar (width times height) is equal to the relative frequency.

It is easy to see that the scale in the figure with 5 ranges has higher values. In order to avoid this change in scale as the number of ranges changes, it is more appropriate to plot the histogram such that the **area** in each **range** reflects the **relative frequency** of the data points within that range. In order to do this for the data in Figure 1.1 we need to do nothing for the 10 ranges case

since the width of each range is 1. For the case of just 5 ranges whose size is 2, we need to divide the relative frequency by 2 to obtain the value of the histogram. In this manner, the area of each bar reflects the relative frequency of the data points, as we expected.

In general, as the number of data points increases, we can approximate the resulting histograms by a continuous curve, where the area under the curve over an interval is equivalent to the relative frequency of the number of data point in the relevant interval. We thus obtain the concept of *probability density*, as shown in Figure 1.2 for the histogram shown in Figure 1.1, where each vertical value in Figure 1.1 was plotted at the mid-point of each range. The results for both 5 ranges and 10 ranges are shown in Figure 1.2 (for the 5 ranges, we had to divide the vertical scale by 2 to preserve the area, as explained above). We see that there is very little difference between the two curves!

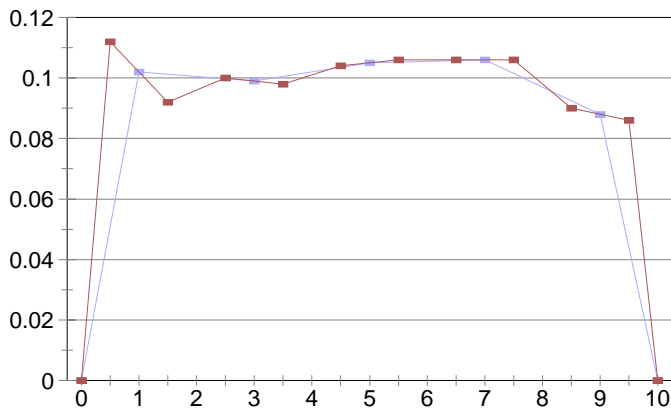


Figure 1.2 Histogram of the same data points plotted so that area under the curve reflects the relative frequency.

What happens when the data has a wide range of values with sparse distribution of data in the highest and lowest ranges? In such cases we may use ranges that are not equal in size. The result will still provide a good histogram provided we use the area under each bar as a measure of the relative frequency.

1.2.2. Data Parameters

Before we can address box plots we need to derive alternative methods of data characterization. Sometimes, a pictorial representation (histogram) conveys too much information, even though it provides a more complete description of the data behavior. For purposes of comparison of different sets of data, we would like to define a few parameters that help us characterize the data. (Usually, when you wish to present data to your boss, he or she expects just a few numbers rather than a complete picture. We all know that bosses have very short attention spans and can handle only a few numbers at a time.)

Several such parameters are defined to characterize a central value around which the data is centered. Another set of parameters characterizes the spread of the data values about this central

value, whether such a spread is large or small. As a measure of center we use two different parameters: *sample mean* and *sample median*. We use the qualifier "sample" to indicate that the results are valid only for the data sample provided. We shall see in later chapters how this "sample" parameter can provide information about the entire population or the entire system.

We can identify several parameters that can describe the properties of the collected data in some manner. Obviously, such parameters cannot convey the information contained in a histogram. However, we do need such parameters in order to compare two histograms as they may look different qualitatively, but we need to express the difference in a quantitative manner. Here we shall define 6 such parameters that can be used to characterize a data set. We sometimes refer to the parameters as belonging to a histogram or to a probability density function.

- a. **Sample Median:** \tilde{x}
- b. **Sample Mean:** \bar{x}
- c. **Sample Standard Deviation:** s
- d. **Quartiles:** Q_1 and Q_3
- e. **Inter-quartile Range:** $IQR = Q_3 - Q_1$
- f. **Percentiles**

9	9	
23	12	
45	12	
32	13	
24	14	
13	15	
42	16	Q1
12	17	
63	17	
14	19	
28	20	
17	23	
12	24	
26	26	median
48	27	
35	28	
27	30	
16	31	
45	32	
15	35	
66	41	Q3
17	42	
31	45	
20	45	
19	48	
30	63	
41	66	

Table 1.1 27 data points.

a. Sample Median:

The median of the sample of data points is the mid-point of the values taken by the data, so that 50% of the data points are above the median and 50% are below the median.

Table 1.1 displays a data set of 27 data points ranging between 9 and 66. The raw data is shown in the left column, and the sorted data is shown in the right column.

The median is denoted by the symbol: \tilde{x} . Note that if high and low values of the data points are changed, the median remains unchanged, if we do not add any new data point. We say that the median is *resistant* to outliers, meaning that it is not susceptible to large deviations in the data. Since economic data (such as income) tends to vary a great deal over a wide range, a good measure of the central point to use is the median. On the other hand, when using grades in a course or exam one need not use the median, as all the grades are between 0 and 100 – so the existence of outliers is not a serious problem.

The values in Table 1.1 are sorted in order to find the middle point, which yields the median value of 26.

Another measure of a central value of the data is the *sample mean*. We can define the sample mean to be the arithmetic average value of the data points.

b. Sample Mean:

Suppose we have n data points which we shall denote by $\{x_i\}$, where $i=1,2,3,\dots,n$. The sample mean is defined as the sum of the values of the data points divided by the number of data points, n , and is shown in Equation (1.1). It is also called sample average, as it is obtained by averaging the values of the data points in the sample.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = (\mathbf{x}_1 + \mathbf{x}_2 + \dots + \mathbf{x}_n)/n \quad (1.1)$$

For the example shown above the sample average is equal to 28.52, which is larger than the median as the values above the median are larger. If we increase some of the higher values or decrease some of the lower values the sample mean will also increase or decrease, respectively. It implies that the sample mean is *not* resistant to outliers. As mentioned earlier, in economic data, such as household income, the median is used rather than the mean, as the mean can be affected by extremely high income of millionaires and billionaires, which may skew the result.

Alternative Evaluation of the Sample Mean:

Suppose we have only a finite number of values obtained in our data set, so that each value appears a number of times. Let the values be denoted by $\{a_i\}$, where $i=1, 2, 3, \dots, N$. Suppose that the value a_i appears n_i times in our sample so that

$$n = n_1 + n_2 + \dots + n_N = \sum_{i=1}^N n_i$$

In computing the average we have to add n_i times each value a_i and then divide the total by n . In other words, instead of adding the value so many times, we can just multiply the value a_i by n_i which is the number of times we have to add it. As a result we obtain the following expression for the sample average:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^N n_i a_i = \sum_{i=1}^N (n_i / n) a_i = \sum_{i=1}^N p_i a_i \quad (1.1a)$$

Here we denoted the relative frequency that the value a_i appears by p_i . We can demonstrate that by looking at a simple example with just six values 1, 2, 3, 4, 5, and 6. There are 150 samples of these six values and their number and relative frequencies are shown in the following table:

a_i	1	2	3	4	5	6
n_i	20	24	29	32	26	19
p_i	0.133	0.160	0.193	0.213	0.173	0.127

We can compute the average directly or by using the formula in (1.1a) and we obtain the same result of 3.513. The histograms (both raw and in terms of relative frequencies) are shown in Figure 1.2a and 1.2b.

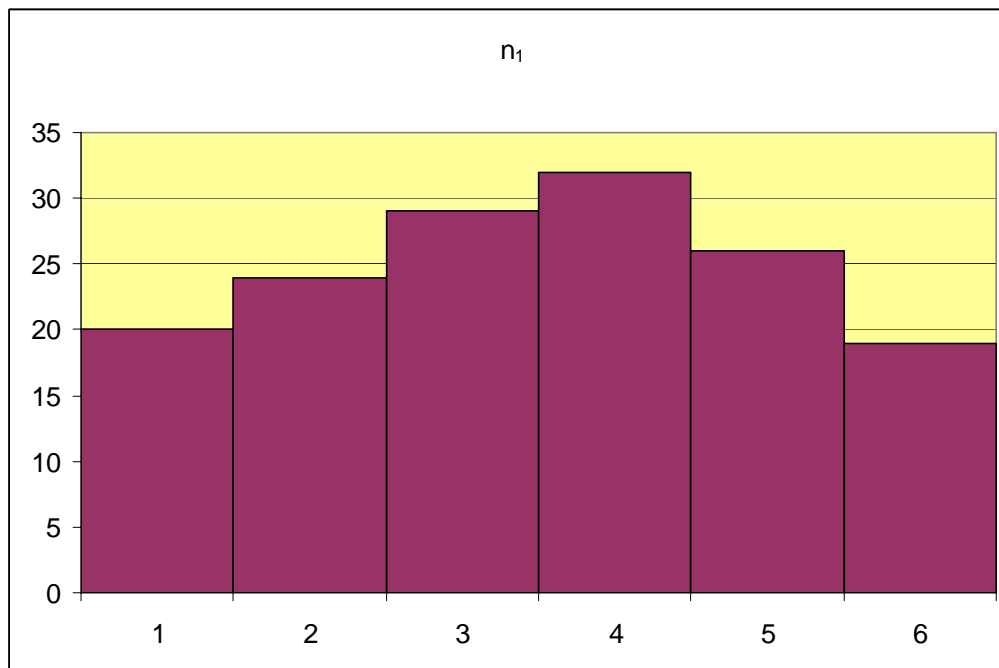


Figure 1.2a. Histogram of the six valued experiment, using raw data

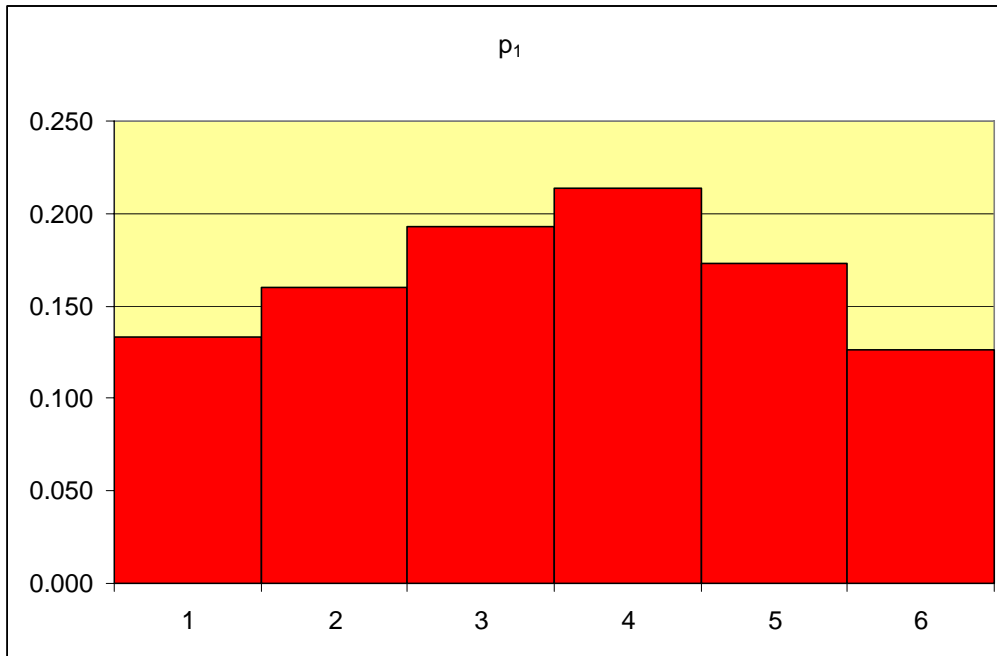


Figure 1.2b. Histogram of the six valued experiment, using relative frequency

Suppose that we now wish to draw the histogram using two values together. This may be needed when we have a large number of possible values. This means that the relative frequency will be about twice as large as when each value occupies one bar in the histogram. We shall draw this in the two ways described earlier, the first using the relative frequency as the values of each bar (Figure 1.2c), and the second, using the relative frequency to represent the area under the histogram (Figure 1.2d).

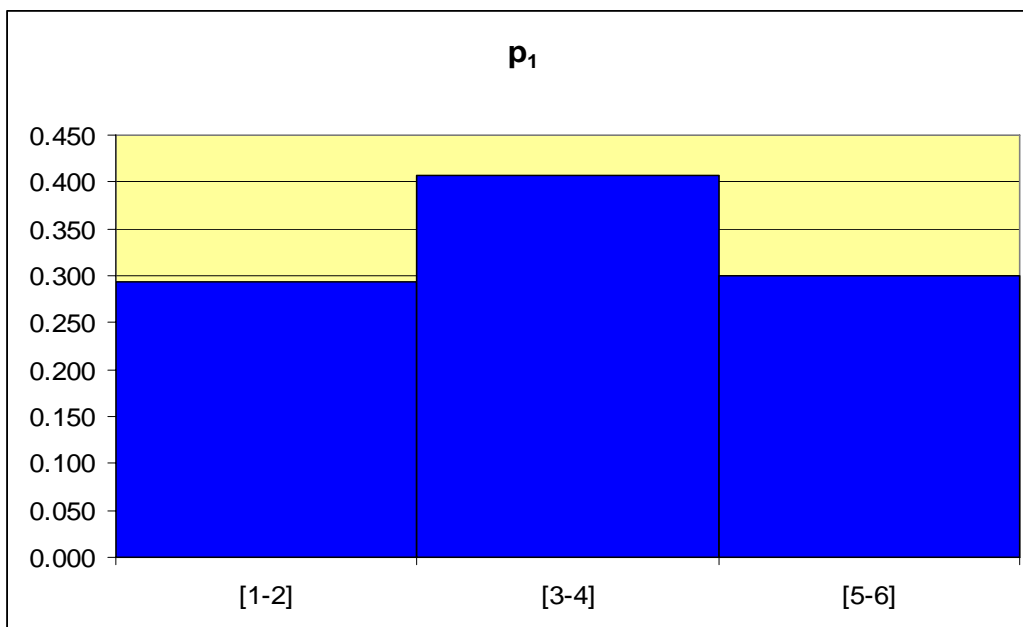


Figure 1.2c. Histogram of the six valued experiment, combining every two values and using relative frequency.

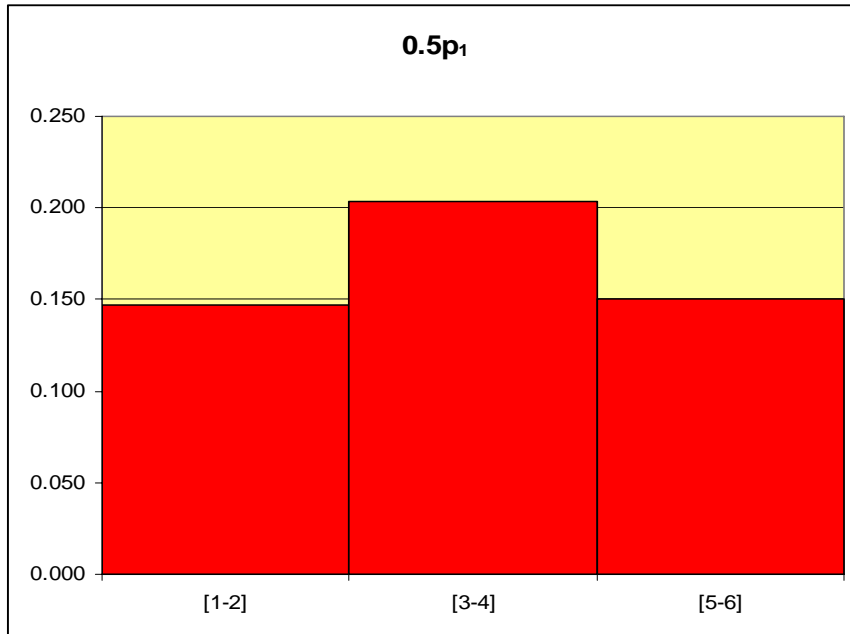


Figure 1.2d. Histogram of the six valued experiment, combining every two values and using area to represent relative frequency.

We now can compute the sample average by multiplying each central value in each bin (1.5, 3.5, and 5.5) by their relative frequency, which is obtained by the area under each bar. The result 3.5133 is an approximation of the correct sample average since we combined every two values.

The mean and the median represent the central value of the data set. This implies that the data values fluctuate around the mean and the median. A measure of the spread of the data values about one of these central values of the mean or the median is needed to provide the information on how far the values deviate from the central value. Such a measure is needed to distinguish between narrowly shaped histograms versus widely shaped histograms. In a narrow histogram, most of the values will be close to the mean or the median. In a wide histogram, the values will spread further from the mean or the median. An example of a widely shaped histogram is one of the temperature in Chicago, while a narrowly shaped histogram is one of the temperature in San Diego..

c. Sample Standard Deviation:

The sample standard deviation, s , is defined as the square root of the **sample variance** s^2 . The sample variance is obtained by the following expression, which measures the average of the squares of the deviations from the sample mean and is shown in Equation (1.2). The reason we use the squares of the values is so that we obtain a positive quantity indicating distance from the sample mean:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (1.2)$$

It can be also expressed as:

$$s^2 = \frac{1}{n-1} \left\{ \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right\} \quad (1.3)$$

Here we also can verify that the standard deviation is **not** resistant to outliers. It also means that if there are many values far from the mean the sample standard deviation may be very large and may approach infinity in some cases.

Why do we divide by $(n-1)$ and not by n ? We do so since we already used the data to obtain the sample mean, so that only $(n-1)$ sample values provide independent information about the spread. You may check to see that the sum of all the deviations from the mean is equal to zero, so these deviations have only $(n-1)$ degrees of freedom!

A measure of the spread that is resistant to outliers is represented by the quartiles.

d. Quartiles:

The quartiles are defined as the points at which 25%, 50%, and 75% of the data points fall. Thus we have the first quartile, Q_1 , which is the value that 25% of the data points fall below. Another definition of the first quartile is that it is the median of the lower 50% of the data points. The second quartile is also equal to the median and represents the value that 50% of the data points fall below or above. The third quartile, Q_3 , is the value that 75% of the data points fall below. Another way of defining the third quartile is that it is the median of the top 50% of the data points. Obviously, quartiles are resistant to outliers. For the data in Table 1.1 we have the following values for the quartiles: $Q_1 = 16$ and $Q_3 = 41$.

e. Inter-Quartile Range (IQR):

A measure of the spread of the data sample, is obtained by the Inter-Quartile Range (IQR), which is defined as

$$IQR = Q_3 - Q_1 \quad (1.4)$$

For the example in Table 1.1 the $IQR = 25$. For the same example we have as standard deviation $s = 15.06$.

f. Percentiles:

A generalization of the quartile concept is to use p^{th} percentiles, where p is a value between 0 and 100. The p^{th} percentile is defined as the point below which $p\%$ of the data points fall. Thus the first quartile is the 25th percentile and the third quartile is the 75th percentile. Obviously, the median is the 50th percentile. It is common to use the 90th and the 10th percentile to indicate the higher and lower values of the data sample. Such a parameter is usually used in ranking students and is usually requested in graduate school or college application forms. It is also used to define

high or low income brackets in the population.

1.2.3. Box Plots:

We can capture the information provided by the median and the quartiles graphically, by using a box plot. A box plot, displays a box indicating the range of values of the data sample within the IQR. The box is bound by the third and first quartile, with a line in the middle to indicate the median. Lines outside the central box extend to the maximum and minimum values of the data. However, if the data extend beyond the box by more than $1.5(IQR)$, then we show points above such lines as outliers. A box plot for the example in Table 1.1 is shown in Figure 1.3. We would have outliers if data points with values higher than $41 + 1.5 \times 25 = 78.5$ were present. Assuming our data is strictly positive, we would not get outlier in this example for too low values of data points.

Some comments about the median and the quartiles: if there is a gap between the lower 50% and the upper 50% of data points, we take the arithmetic average of the two points in the gap between the upper and lower 50% to compute the median. We handle the computation of the quartiles in a similar manner, so that if there is a gap where the quartile is to occur, we average the data points at this gap to evaluate the quartile.

Another measure of the spread of the data point is called the **range**, which is simply the difference between the highest and lowest data point. It usually is of little interest in normal situations, since such a measure is strongly susceptible to outliers. For examples, in computing statistics on income in a population, one billionaire could totally skew the data and hence the range is a meaningless measure.

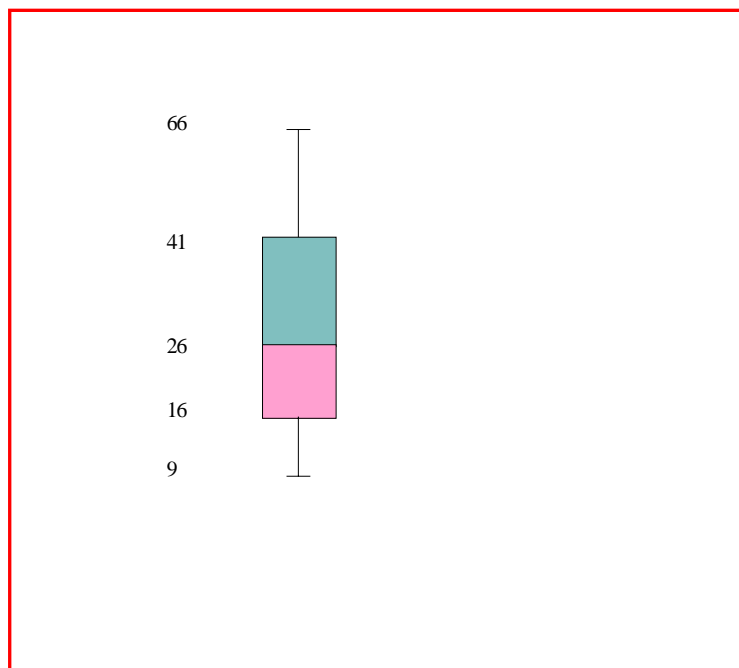


Figure 1.3 Box Plot for the data sample in Table 1.1

1.2.4. Examples

Example 1.1:

Consider the data collected about errors in a binary communications channel. We can just record the number of errors and their locations to provide a complete picture of the data. If we count the number of errors in n bits as m , then we can identify a data point as “one” if the bit is in error, and as “zero” if the bit is received correctly. Thus there are m “ones” and $(n-m)$ “zeros” in the data collected. We therefore can obtain the sample mean, \bar{x} , and the sample variance, s^2 , of the errors in the channel by using the following expressions:

$$\bar{x} = \frac{0(n-m) + 1m}{n} = \frac{m}{n}$$

$$s^2 = \frac{(1-\bar{x})^2 m + (0-\bar{x})^2 (n-m)}{n-1} = \bar{x}(1-\bar{x})\left(\frac{n}{n-1}\right)$$

The sample mean is also equal to the probability of an error in the channel and is also called the *error rate*. The sample variance for large values of n becomes just $s^2 = \bar{x}(1-\bar{x})$, which only depends on the error rate. It is difficult to speak in such a case about the median and other parameters as they do not make sense for this example.

Example 1.2:

Given in Table 1.2 are statistics obtained from a voice digitizing system, where the length of calls in number of packets are recorded. The table shows both the number and the relative frequency.

Length k	1	2	3	4	5	6	7	8	9	10	Totals
Number	105	75	45	30	21	12	6	3	2	1	300
P= Relative Frequency	0.35	0.25	0.15	0.1	0.07	0.04	0.02	0.01	0.0067	0.0033	1
Approx.	0.35	0.23	0.15	0.096	0.062	0.041	0.026	0.0172	0.0111	0.00725	0.99055

Table 1.2 Data on call lengths in packets, corresponding relative frequency and approximate model

The median of the length of call is 2 packets, while the sample mean is 2.55. The histogram is shown in Figure 1.4. It is seen that the histogram may be approximated by an exponential curve which by careful curve fitting is selected as $0.35(0.65)^{k-1}$, for a packet of length k . The approximation is also shown in the table and in the Figure.

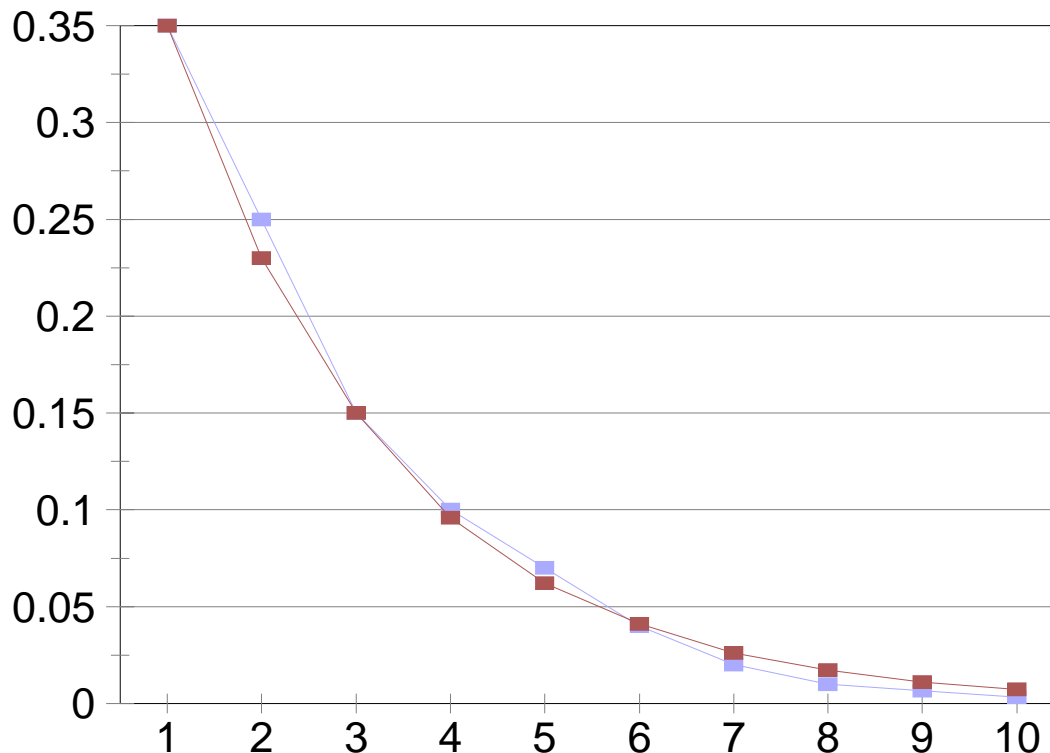


Figure 1.4 Comparison of the call length data in Example 1.2
The blue curve represents the data and the red curve is the approximating model.

Example 1.3:

Given in Table 1.3 a histogram obtained from the first 600 messages in my e-mailbox of November 2006, where the length of messages in KBytes is shown. The table shows both the number (N) and the relative frequency (P) for message length k. (Note that the value k=6, for example, represent bins between 4 and 6 KBytes.)

k	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	Total
N	213	134	58	60	28	21	27	14	10	4	3	3	5	6	3	589
P	0.355	0.223	0.097	0.1	0.047	0.035	0.045	0.023	0.017	0.007	0.005	0.005	0.008	0.01	0.005	0.982
Model	0.35	0.227	0.148	0.096	0.062	0.041	0.026	0.017	0.011	0.007	0.005	0.003	0.002	0.001	0.001	0.997

Table 1.3 Data on message lengths in KBytes, corresponding relative frequency and approximate model

The histogram is shown in Figure 1.5. It is seen that the histogram maybe approximated by an exponential curve which by careful curve fitting is selected as $0.35(0.65)^{0.5k-1}$, for a message of length k. The approximation is also shown in the table and in the Figure. It is interesting to see that the model drops off sharply, while the actual data has longer tails. We shall address this

longer tail issue at a later chapter.

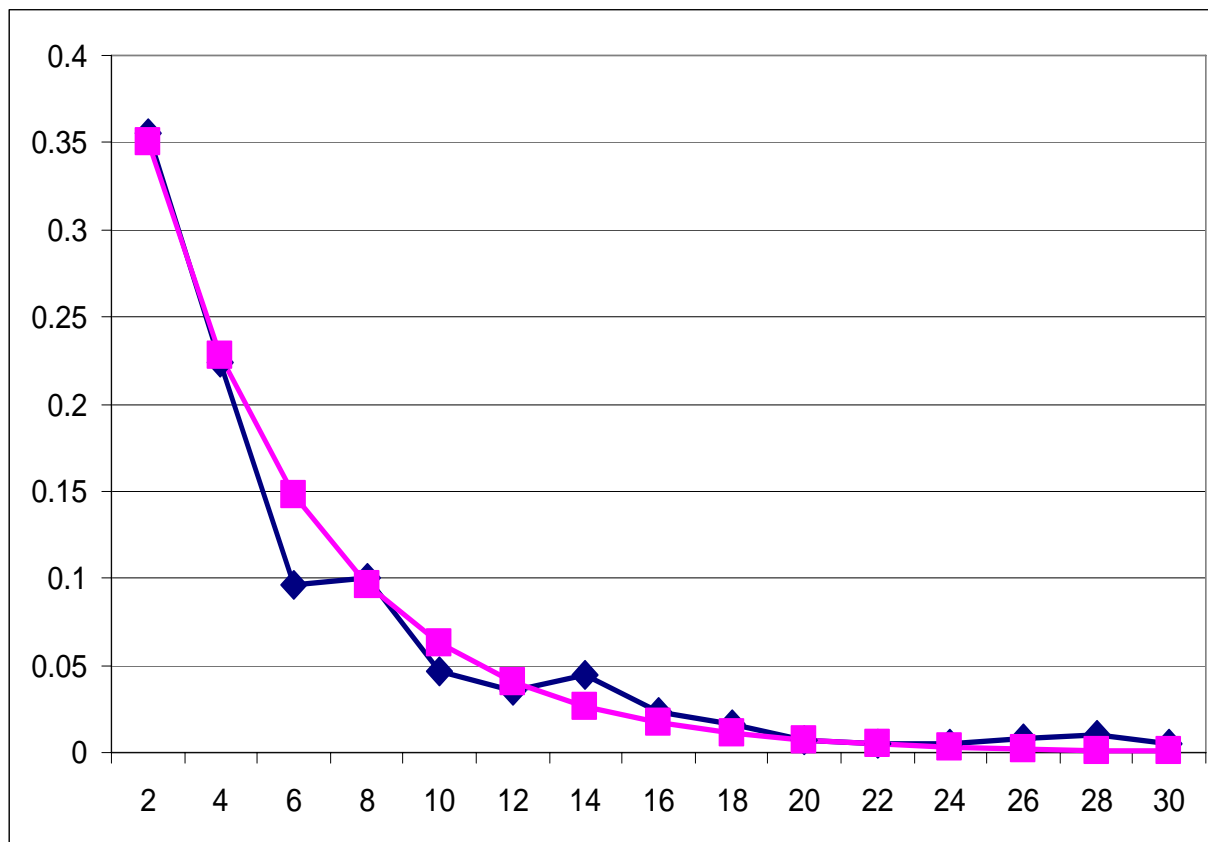


Figure 1.5 Comparison of the message length data in Example 1.3
The blue curve represents the data and the pink curve is the approximating model.

Example 1.4:

In this example 100 random numbers were obtained by spinning a pointer and measuring its extended point of impact. The numbers are shown in Table 1.4.

The histogram is shown in Figure 1.6. It shows that there are many outliers which imply that the standard deviation is bound to be very large. In the Figure we see the rise in the two extreme ranges as they represent any data point higher than 5.5 or lower than -5.5 . This observation is confirmed by the numerical values. We see that the minimum value obtained is -49.39 and the maximum value obtained is 120.13 . The range of the data is therefore 169.52 . The mean is 1.53 (it should be closer to zero, except that we find here more positive outliers than negative ones). The standard deviation is 16.76 , which is high due to the wide range of the data. We now compare these values to the median and quartiles. The median is -0.05 , as expected, and the two quartiles are: $Q1 = -1.59$ and $Q3 = 1.15$, so that the $IQR = 2.74$ is much smaller than the standard deviation.

-0.07	-0.39	-4.41	-49.39	-1.25	-2.02	-2.51	-4.75	-5.08	1.65
-2.72	-0.51	0.48	0.12	-0.46	0.04	0.33	-4.88	3.43	1.06
18.69	-1.62	-0.17	1.14	-4.64	-0.91	120.13	3.05	0.33	-1.44
-2.68	-0.76	0.66	-1.59	-0.64	0.18	0.03	-0.61	0.38	-6.11
-2.19	-1.05	-7.41	2.95	-0.99	-0.51	1.79	1.14	0.65	-1.48
-8.22	-1.12	1.26	-1.59	3.80	2.04	-1.86	2.41	-0.04	0.11
-0.55	0.49	3.09	-0.93	0.50	68.41	2.94	-4.89	-0.10	64.20
2.59	-1.99	0.08	0.24	7.18	-0.42	-29.45	1.10	-0.50	-8.35
0.81	0.74	-1.11	1.81	1.17	-0.42	8.21	3.60	-2.16	2.76
1.99	-17.82	11.64	1.48	-2.65	-0.03	-0.03	-2.56	-0.76	1.14

Table 1.4 A sample of 100 numbers representing the impact point of a random spinning pointer.

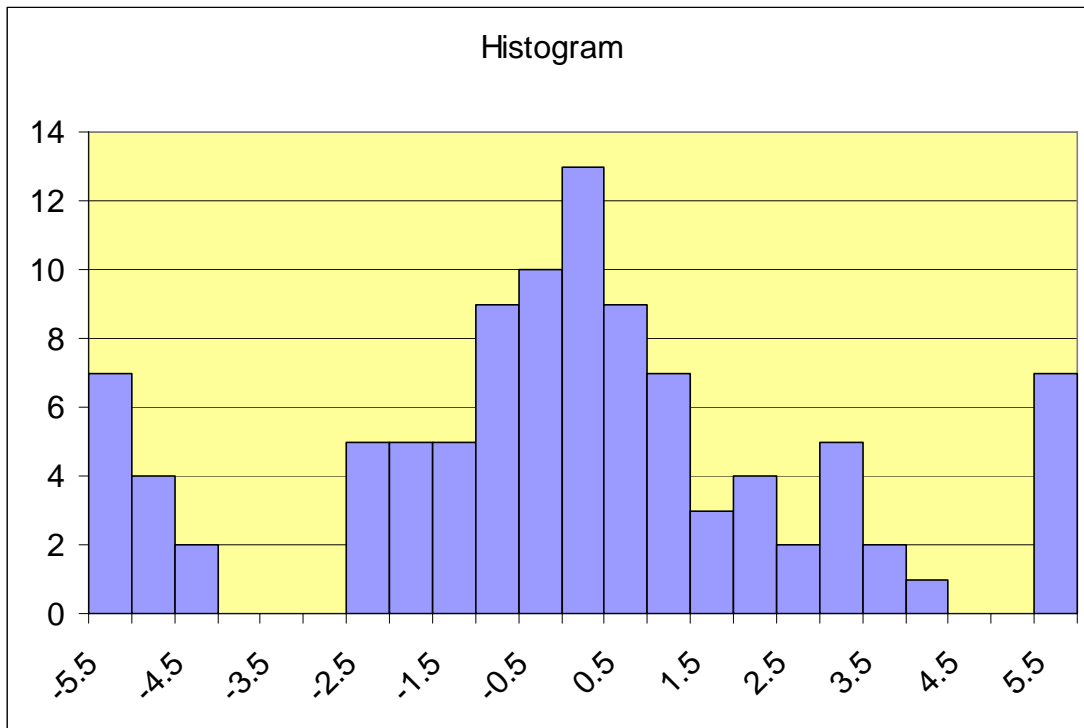


Figure 1.6 Histogram for the data in Table 1.4

2.0 Probability Models

2.1 Introduction to Probability

Probability models provide a more systematic approach to the analysis of data collected about information and telecom systems as well as many other man-made systems. The models provide also a means for system design that will satisfy quality of service criteria that could not be accomplished without using statistical methods. It also provides tools for evaluating systems for acquisition or implementation. Many of these models also apply to non-engineering systems by exploiting special rules and properties that address insurance, finance, and other well-defined problems. In order to discuss probability models we need to define the appropriate setting, namely, the concept of a *random experiment*. Usually, when we perform an experiment, we can predict the answer in advance. For example, if we throw a ball in the air, we can plot its trajectory with very good precision. When we cannot determine the outcome of the experiment in advance, we need to rely on what we shall define as a random experiment.

A Random Experiment: A random experiment is an experiment with more than one possible outcome. It is not known in advance which outcome will occur.

An Event: An event is a collection of outcomes of the random experiment under consideration. The collection maybe empty, may contain just one outcome, or may contain some or all outcomes.

Probability: Probability is a measure indicating the likelihood of an outcome (or collections of outcomes, i.e. **events**) to occur in an **experiment**, which may have more than one possible outcome (a **random experiment**).

Hence, **probability** is assigned to **events** occurring in a **random experiment**, which is an experiment with more than one possible outcome.

Examples of random experiments are:

- We transmit a bit: it can be received with error or received correctly,
- We transmit a packet: it is either dropped or delivered properly,
- We transmit a packet: it can be delivered with various possible delays,
- At any instant a network node or branch may be operating or failing,
- We transmit a message: its lengths may have different values,
- We observe a phone call: it may last a different amount of time,
- A wireless call in progress: it may be dropped or completed successfully,
- We dial for an outside line: we either get a line or get a busy tone,
- We observe messages arriving at a switch: arrival times and arrival rates may vary,
- We observe a credit card company in a day: number of fraudulent charges may vary,
- We observe a traffic intersection: the number of accidents in a month may be different,
- We observe a disk drive for one year: it may or may not crash,
- We observe the price of a given stock: it may go up or down by different amounts.

As we see above there is practically limitless ways of defining random experiments.

2.1.1. Random Experiment and Events Algebra

We define the **sample space** (and denote it by S) as the collection of **all** possible outcomes in a random experiment. As an example in rolling two dice, we have 36 possible outcomes reflecting the different pairs of dots on the two dice. In a system with 5 components, with each component operating properly or failing, we have 32 possible outcomes reflecting the various combinations involving failures of one or more components. In observing a byte (8 bits) of data, it may have 256 different possible outcomes depending on the “ones” and “zeros” in each of its bits. We observe a phone call, it may last anywhere between zero and infinity seconds (the latter value occurs if it involves teenagers). In this case any real positive value is a possible outcome. Events would then be any combination of these values, such as a call lasts more than 2 minutes, or a call lasts between 1 and 3 minutes, or a call lasts less than 3 minutes.

We define **events** as collections of outcomes from the sample space S . For example, in the failure of components problem, an event maybe two of the five components failed. Another example maybe defined as the failure of components #2 and #4. We can enumerate similar such events for the dice rolling problem, such as the rolls are the same, the sum of the rolls is even, or the sum of the rolls is less than 5.

In general we describe the sample space and the events it contains as a rectangle with events shown as closed areas inside the rectangle. Events are denoted by the letters A, B, C , etc. The resulting graphical representation is called a **Venn diagram**.

When we describe events in general terms we usually draw them as ellipses within the rectangle S , representing the sample space. When dealing with a specific example we draw the events as rectangles and usually we draw them such that the area of the rectangle representing an event is proportional to the probability of that event.

We next consider some types of events and how we can define operations involving events. Since events are logical entities, the operations are logical and not numeric operations.

The **Null Event**:

The null event is the event that contains **no** outcomes and is denoted by \emptyset . The null event is also called the **impossible event**, since it cannot happen.

The **Certain Event**:

The certain event, S , is the event that includes **all** outcomes, and is the same as the entire sample space. Therefore it must happen as one of its outcomes is bound to occur.

Example 2.1:

If we roll two dice one red and one green and define

$$A_i = \{\text{the green die rolled } i \text{ dots}\}$$

$$B_j = \{\text{the red die rolled } j \text{ dots}\}$$

The certain event S contains 36 individual outcomes (A_i, B_j) for $i, j = 1, 2, 3, 4, 5, 6$. An example of a null event is the following: $\{A_5 \text{ and the sum is less than } 4\}$.

In order to be able to assign probability to events and some operations involving events we need to describe operations involving events and thus we obtain an algebra of events. These include the following logical operations. These also have mathematical symbols, but we prefer to use the more descriptive terms, rather than the abstract math symbols (called unions and intersections).

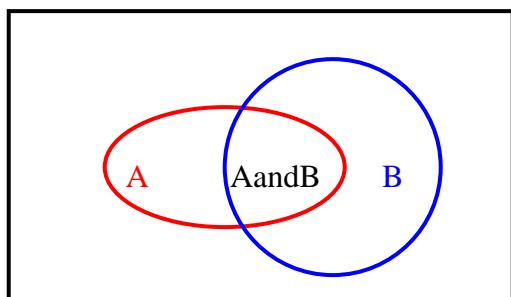


Figure 2.1b. The event (A **and** B)

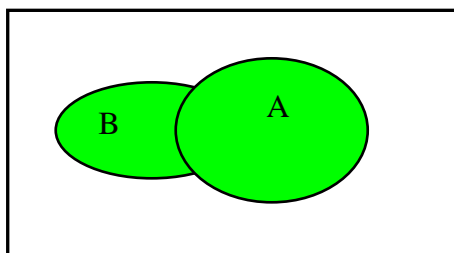


Figure 2.1a. The event (A **or** B)

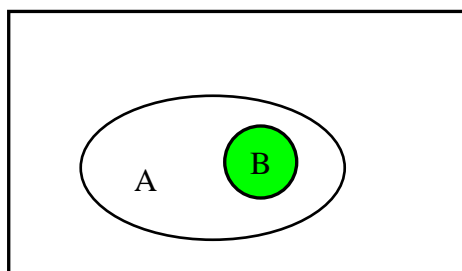


Figure 2.1d. A contains B

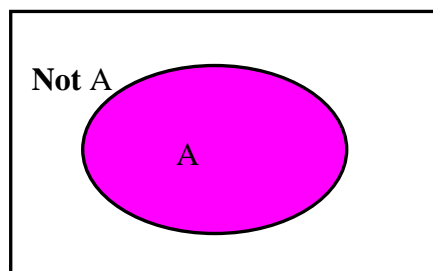


Figure 2.1c. Complement: **not A**

OR:

The “**or**” operation involving two or more events is the event that includes the outcomes that are in either of the events or in both. It is sometimes called “**inclusive or**” to indicate that it also contains the outcomes that are in both. In other words:

$$(A \text{ or } B) = \{\text{Outcomes that are in either A or in B or in both}\}$$

This is shown graphically in Figure 2.1a. In our dice example the event

$$(A_3 \text{ or } B_3) = \{\text{that we rolled a “three”}\}.$$

AND:

The “**and**” operation involving two or more events is the event that includes the outcomes that are included in both events. In other words:

$$(A \text{ and } B) = \{\text{Outcomes that are in both A and B}\}$$

This is shown graphically in Figure 2.1b. In the dice rolling example, the event

$$(A_3 \text{ and } B_3) = \{\text{that both dice rolled a “three”}\}.$$

NOT:

NOT, which is also called the *complement* of an event A, is the event that A does not happen, and is denoted by A^c . In other words,

$$A^c = \{\text{All outcomes that are not in A}\}.$$

This is shown graphically in Figure 2.1c. Clearly, the null event and the certain event are complements of each other.

We note that we have the following properties:

$$S = (A \text{ OR } A^c)$$

$$\emptyset = (A \text{ AND } A^c)$$

Disjoint or mutually exclusive events:

Events that do not contain any elements in common are called **disjoint** or **mutually exclusive**. In other words, mutually exclusive events satisfy the relation:

$$(A \text{ and } B) = \emptyset$$

Another way to describe mutually exclusive events is to state that they cannot occur together. If one of these events occurs then the other does not occur. Graphically, it means that they are represented by two geometrical shapes with no overlapping region.

Contains:

Finally, an event B is **contained in** A (or A **contains** B) if all the outcomes in B are also outcomes of A. This is shown graphically in Figure 2.1d. For example, in rolling two dice, if we define A as the events that the sum of the rolls is not larger than 7, and B as the events that one of the rolls is 1, then A contains B.

We can relate the operations shown above by the following properties, which will not be proved here. You can verify these properties by using Venn diagrams.

The order of the events involved in **and** or **or** operations is not important (i.e. the operations commute):

$$\begin{aligned}(A \text{ or } B) &= (B \text{ or } A) \\ (A \text{ and } B) &= (B \text{ and } A)\end{aligned}$$

When more than two events are involved, it does not matter which we perform first:

$$\begin{aligned}((A \text{ or } B) \text{ or } C) &= (A \text{ or } (B \text{ or } C)) = (A \text{ or } B \text{ or } C) \\ ((A \text{ and } B) \text{ and } C) &= (A \text{ and } (B \text{ and } C)) = (A \text{ and } B \text{ and } C)\end{aligned}$$

The following properties are called distributive properties and may be familiar in multiplication relative to addition: $a \times (b+c) = a \times b + a \times c$:

$$\begin{aligned}A \text{ and } (B \text{ or } C) &= (A \text{ and } B) \text{ or } (A \text{ and } C) \\ A \text{ or } (B \text{ and } C) &= (A \text{ or } B) \text{ and } (A \text{ or } C)\end{aligned}$$

Finally, the last property deals with complementation:

$$\begin{aligned}\text{NOT } (A \text{ or } B) &= (\text{Not } A) \text{ and } (\text{Not } B) \\ \text{NOT } (A \text{ and } B) &= (\text{Not } A) \text{ or } (\text{Not } B)\end{aligned}$$

As an example of the last set of equation consider the tossing of three coins:

Let $H_i = \{\text{Coin \#}i \text{ is Heads}\}$, and $T_i = \{\text{Coin \#}i \text{ is Tails}\} = H_i^c$, for $i = 1, 2, 3$.

Define the event $A = \text{at least one Heads} = \{H_1 \text{ or } H_2 \text{ or } H_3\}$. Its complement is equal to the event $A^c = \{\text{No Heads}\} = \{T_1 \text{ and } T_2 \text{ and } T_3\} = \{H_1^c \text{ and } H_2^c \text{ and } H_3^c\}$. We shall observe many other examples of more interest in the sequel.

2.1.2 Definition of Probability

We are now ready to define what constitutes a valid probability assignment to events of a random experiment. There are three basic approaches to assign probability to events:

- a. Geometric or logical approach

- b. Experimental approach using data
- c. Subjective approach
 - a. The most logical is the geometric approach, which is based on an analysis of how the experiment is conducted. It is usually based on some physical consideration of the way the experiment is designed. It applies to some natural systems as well as many man-made systems.
 - b. In many cases involving economic and social events, we have to rely on the experimental approach, by using data collected on the events of interest. In many biomedical and other health related systems, we have to depend on extensive data collection in order to obtain a valid probabilistic model or result. Even in engineering systems, we are dependent on data collection to validate a model or obtain the numerical values of its parameters.
 - c. Finally, there is the subjective approach, which we shall not address here, used by gamblers and lottery players (and even by stock market investors). It is based on some metaphysical or superstitious beliefs about what is expected to happen. It is even used by students who took extensive courses in probability, when they are not in a classroom setting.

We shall work with the geometric approach to define some valid probability models. In general the resulting models will have parameters that need to be assigned values. We then use data simply to identify the values of the parameters of interest and the validity of the model.

Since we can assign probability by using logical or geometric arguments or from collected data, one may ask if there are any rules for such definitions. Indeed, no matter how we assign probability, in each case the assigned probability must satisfy certain rules, which are called the axioms of probability. These rules cannot be proved or disproved, as they simply represent a convention for defining valid probability models.

Axioms of Probability:

Suppose we are given a random experiment, with sample space S , and a collection of events. To each event A in S we assign probability $P(A)$ such that it satisfies the following three axioms:

- 1. $P(A) \geq 0$**
- 2. $P(S) = 1$**
- 3. If A and B are mutually exclusive, then $P(A \text{ or } B) = P(A) + P(B)$**

These axioms or rules specify that probability is not negative, it is defined as 1 for the certain event, and when we assign probability to two events that cannot occur together, the probability that either can occur is obtained by adding their probabilities.

Based on these axioms we can obtain several more rules that are satisfied by any probability assignment. These can be derived from the axioms, but the derivations will not be presented in these notes. Most of the properties should be intuitively understood.

Additional Rules of Probability:

4. $P(A^c) = 1 - P(A)$

5. $P(\emptyset) = 0$

6. $P(A) \leq 1$

7. If A contains B then $P(A) \geq P(B)$

8. For any A, B we also have: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

9. $P(A \text{ or } B) = P(A) + P(B \text{ and } A^c) = P(B) + P(A \text{ and } B^c)$

The fourth rule shows that if A occurs with 40% probability, then its complement will occur with probability of 60%. Rule five simply states that the null event has zero probability. It never happens! However not every event with zero probability cannot occur. For example, to measure a variable to be exactly 1.5 may have probability zero, while the probability that the variable takes values between 1.499 and 1.501 is not zero. Hence, 1.500000 (with infinitely many zeros following the 5) can occur, but exactly that value may have probability zero. Rule 6 states that a probability cannot exceed 1. If it does in any of your work, it means that you made a big mistake, so check your work.

How to assign probabilities to cases where we have finitely many outcomes?

If we can divide the sample space into a collection of n mutually exclusive events, then all we need to do is assign a probability of p_i to each of these events, such that the total is equal to 1. If these events happen to be equally likely, then each will be assigned a probability of $1/n$.

A good example to such an assignment is the dice rolling case discussed in Example 2.1. Here we have 36 mutually exclusive events, and there is no reason to prefer one outcome over another, hence each will have a probability of $1/36$. We can use the resulting assignment to solve problems involving more complex events using the rules described above. As an example, consider the last rule (rule 8) to solve for the probability that one of the dice rolls a 3:

$$P(A_3 \text{ or } B_3) = P(A_3) + P(B_3) - P(A_3 \text{ and } B_3) = 1/6 + 1/6 - 1/36 = 11/36$$

It should be noted that if we just add the two probabilities we would obtain a larger value than the correct probability. We would obtain $1/3$ instead of $11/36$. By just adding the two probabilities we would have included the probability that both were “3” twice. We could also obtain this probability directly by counting how many individual events show a “3” on one of the rolls, and that is exactly 11 out of the 36 total events.

Example 2.2:

Suppose that in a packet-switched network we observe that 1% of all packets are dropped or not delivered properly. We can define a random experiment involving the sending of a single packet with two events:

$$A = \{\text{packet is received properly}\}$$

$$A^c = \{\text{packet is NOT received properly}\}$$

Lacking any other information we may therefore assign the following probabilities to these two events:

$$P(A) = 0.99$$

$$P(A^c) = 0.01$$

If we now try to extend this definition to the transmission of more than two packets, we find that we need additional rules or properties to be able to do so properly. If on the other hand instead of packets we tossed two coins we could answer such questions the same way we handled the dice-rolling example, by dividing the sample space into four equally likely events so that each of these events will have probability of 0.25 (assuming the coins are fair).

We therefore need to introduce a new concept that helps us figure probabilities of events defined by the “**and**” operation of two or more events. We arrive here at the concept of *independent* events.