Basics of Statistics

Statistics is a very broad subject, with applications in a vast number of different fields. In generally one can say that statistics is the methodology for collecting, analyzing, interpreting and drawing conclusions from information. In other words, statistics is the methodology which scientists and mathematicians have developed for interpreting and drawing conclusions from collected data. Statistical methods can be used to find answers to the questions like:

- What kind and how much data need to be collected?
- How should we organize and summarize the data?
- How can we analyse the data and draw conclusions from it?
- How can we assess the strength of the conclusions and evaluate their uncertainty?

Types of Data

When working with statistics, it's important to recognize the different types of data:

- Quantitative data (numerical)
- Qualitative data (categorical)

Quantitative data can be classified into two types which are discrete and continuous. Discrete data represents items that can be counted where there is no in between values such as the numbers of children in family, the numbers of car accident on the certain road on different days, or the numbers of students taking basics of statistics course. Continuous data represents measurements; their values have increments in between such as length, weight, or temperature.

Qualitative data represents characteristics such as a person's gender, marital status, hometown, or the types of movies they like. Categorical data can take on numerical values (such as "1" indicating male and "2" indicating female), but those numbers don't have mathematical meaning.

Ordinal data mixes numerical and categorical data. The data fall into categories, but the numbers placed on the categories have meaning. For example, rating a restaurant on a scale from 0 (lowest) to 4 (highest) stars gives ordinal data. Ordinal data are often treated as categorical, where the groups are ordered when graphs and charts are made. However, unlike categorical data, the numbers do have mathematical meaning.

For example, if you survey 100 people and ask them to rate a restaurant on a scale from 0 to 4, taking the average of the 100 responses will have meaning. This would not be the case with categorical data.

Measures of Centre

Descriptive measures that indicate where the center or the most typical value of the variable lies in collected set of measurements are called measures of center. In other words, they all show where the center of a set of data ”tends” to be. Each one is useful at different times. Measures of center are often referred to as averages.

The median and the mean apply only to quantitative data, whereas the mode can be used with either quantitative or qualitative data.

Mean

The **mean**, often called the ‘average’ of a numerical set of data, is the sum of all of the numbers divided by the number of values in the data set.

$$\text{Mean} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \sum_{i=1}^{n} x_i,$$

This value is the arithmetic mean, and it tells us what value we would have if all of the data were the same. The mean is a summary statistic that gives you a description of the entire data set and is especially useful with large data sets where you might not have the time to examine every single value. However, the mean is affected by extreme values, called outliers, and can end up leaving the observer with the wrong impression of a data set.

Median

The **median** is the number in the middle position once the data has been organized. Organized data is simply the numbers arranged from smallest to largest or from largest to smallest. This is the only number for which there are as many above it as below it in the set of organized data, and is referred to as the *equal areas point*. The median, for an odd number of data, is the value that is exactly in the middle of the ordered list, it divides the data into two halves. The median for an even number of data, is the mean of the two values in the middle of the ordered list. The median is useful when there are a few extreme values that can affect the mean, because the middle number will stay in the middle. The median often gives a good impression of the center, because there are 50% of the values above the median, 50% of the values below the median, and it doesn't matter how big the biggest values are or how small the smallest values are.

Mode

The **mode** of a set of data is simply the number that appears most frequently in the set. There are no calculations required to find the mode of a data set. You simply need to look for it. However, be aware that it is common for a set of data to have no mode, one mode, two modes or more than two modes. If there is more than one mode, simply list them all. And, if there is no mode, write 'no mode'. No matter how many modes, the same set of data will have only one mean and only one median. The mode is a measure of central tendency that is simple to locate but is not used much in practical applications. It is the only one of these three values that can be for either categorical or numerical data.

Measure of Variation

In addition to locating the center of the observed values of the variable in the data, another important aspect of a descriptive study of the variable is numerically measuring the extent of variation around the center. Two data sets of the same variable may exhibit similar positions of center but may be remarkably different with respect to variability.

Just as there are several different measures of center, there are also several different measures of variation. In this section, we will examine three of the most frequently used measures of variation; the range, the interquartile range and the standard deviation. Measures of variation are used mostly only for quantitative variables.

Range

The range of a data set describes how spread out the data is. To calculate the **range**, subtract the smallest value from the largest value (maximum value – minimum value = range). This value provides information about a data set that we cannot see from only the mean, median, or mode.

For example, two students may both have a quiz average of 75%, but one of them may have scores ranging from 70% to 82% while the other may have scores ranging from 24% to 90%. In a case such as this, the mean would make the students appear to be achieving at the same level, when in reality one of them is much more consistent than the other.

However, in using the range, a great deal of information is ignored, that is, only the largest and smallest values of the variable are considered; the other observed values are disregarded.

Interquartile Range

Before we can define the sample interquartile range, we have to first define the percentiles, the deciles and the quartiles of the variable in a data set. The percentiles of the variable divide observed values into hundredths, or 100 equal parts. Roughly speaking, the first percentile, P1, is the number that divides the bottom 1% of the observed values from the top 99%; second percentile, P2, is the number that divides the bottom 2% of the observed values from the top 98%; and so forth. The median is the 50th percentile.

The deciles of the variable divide the observed values into tenths, or 10 equal parts. The variable has nine deciles, denoted by D1,D2, . . . ,D9. The first decile D1 is 10th percentile, the second decile D2 is the 20th percentile, and so forth.

The most commonly used percentiles are quartiles. The quartiles of the variable divide the observed values into quarters, or 4 equal parts. The variable has three quartiles, denoted by Q1,Q2 and Q3. Roughly speaking, the first quartile, Q1, is the number that divides the bottom 25% of the observed values from the top 75%; second quartile, Q2, is the median, which is the number that divides the bottom 50% of the observed values from the top 50%; and the third quartile, Q3, is the number that divides the bottom 75% of the observed values from the top 25%.

To understand the quartiles, Let n denote the number of observations in a data set. Arrange the observed values of variable in a data in increasing order.

1. The first quartile Q1 is at position n+1/4 ,
2. The second quartile Q2 (the median) is at position n+1/2 ,
3. The third quartile Q3 is at position 3(n+1)/4 ,

in the ordered list.

Next we define the interquartile range. Since the interquartile range is defined using quartiles, it is preferred measure of variation (spread) when the median is used as the measure of center.

The interquartile range of the variable, denoted IQR, is the difference between the first and third quartiles of the variable, that is,

IQR = Q3 − Q1.

Roughly speaking, the IQR gives the range of the middle 50% of the observed values.

Standard Deviation

Another measure of spread that is used in statistics is called the standard deviation. The **standard deviation** measures the spread around the mean. This value is more difficult to calculate than range or IQR, but the formula used takes all of the data values in the distribution into account. Standard deviation is the appropriate measure of variation (spread) when the mean is the measure of center. The symbol for standard deviation of a sample is ***s*** (on the graphing calculators it is ***Sx***).

The standard deviation can be any number zero or greater. It will only be equal to zero if there is no spread (i.e. all values are exactly the same). The more spread out the data is, the larger the standard deviation will be. The standard deviation is most appropriate when you have a very symmetrical, bell-shaped distribution called a normal distribution. We will study this type of distribution in the next section.

In order to calculate the standard deviation you must have all of the values. Then you follow these steps:

1. Calculate the mean of the values.
2. Subtract the mean from each data value. These are the individual deviations.
3. Each of these deviations is squared.
4. All of the squared deviations are added up.

5. This total of the squared deviations is divided by one less than the number of deviations. This is the variance.

6. Take the square root of the variance. This is the standard deviation.

The formula for calculating the variance is:

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

The formula for calculating standard deviation is:

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

However, this formula is very time consuming when you have a large set of data. Also, it is easy to make a mistake in your calculations. We will show the process with a small set of data, but generally we will use our calculator to find the standard deviation.

Outliers

An **outlier** is a value that does not fit with the rest of the data. Some distributions will have several outliers, while others will not have any. We should always look for outliers because they can affect many of our statistics. Also, sometimes an outlier is actually an error that needs to be corrected. If you have ever ’bombed’ one test in a class, you probably discovered that it had a big impact on your overall average in that class. This is because the mean will be affected by an outlier-it will be pulled toward it. This is another reason why we should be sure to look at the data, not just look at the statistics about the data. When an outlier is part of the data and we do not realize it, we can be misled by the mean to believe that the numbers are higher or lower than they really are.

An observation is a suspected outlier if it falls more than 1.5 × IQR above the third quartile or below the first quartile. Mathematically, x is a suspected outlier if

$$x < Q_1 - 1.5 * IQR \text{ or } x > Q_3 + 1.5 * IQR$$

Histogram

When it is not necessary to show every value the way a stem plot would do, a histogram is a useful graph. Histograms organize numerical data into ranges, but do not show the actual values. The **histogram** is a summary graph showing how many of the data points falling within various ranges. Even though a histogram looks similar to a bar graph, it is not the same. Histograms are for numerical data and each ’bar’ covers a range of values. Each of these ’bars’ is called a **class** or **bin.** Histograms are a great way to see the shape of a distribution and can be used even when working with a large set of data.

The width of the bins is the most important decision when constructing a histogram. The bins need to be of consistent width (i.e. all cover a range of 10, or 25, etc.). It is generally a good idea to try to have 7 to 15 bins. Start with the range and divide by 10. This will give you a rough idea of how wide to make your bins. From there it becomes a judgment call as to what is a reasonable bin width. For example, it really does not make any sense to count by 11.24 just because that is what the range divided by 10 is equal to. In such a case, it might make more sense to count by 10’s or 12’s depending on the specific data.

In order to understand the concept of construction the histogram, let's examine the following example. Suppose that the test scores of 27 students were recorded. The scores were: 8, 12,17, 22, 24, 28, 31, 37, 37, 39, 40, 42, 43, 47, 48, 51, 57, 58, 59, 60, 65, 65, 74, 75, 84, 88, 91.

The lowest score was an 8 and the highest was a 91. Now, let's follow the following steps to construct the histogram.

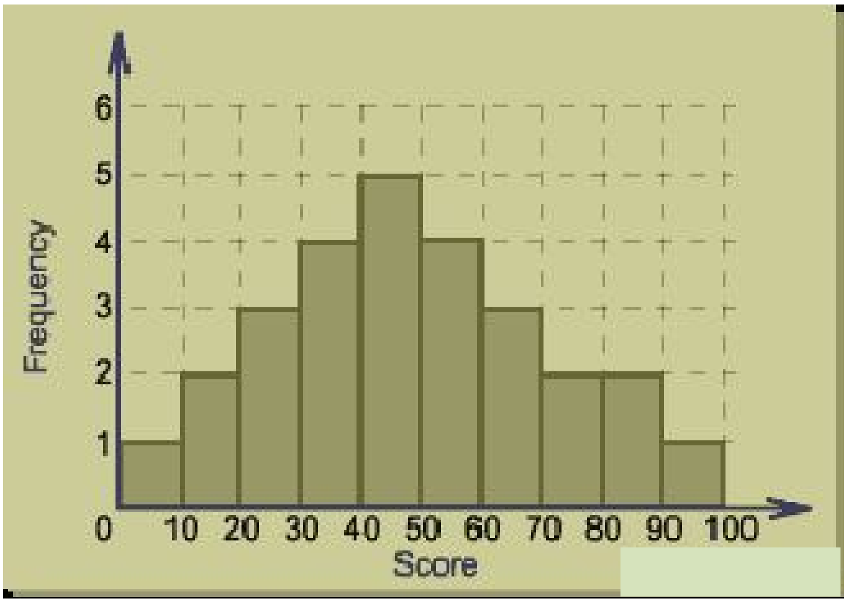
Plan bin width: The first step is to look at the range (91 - 8 = 83). Then divide the range by 10 (83/10 = 8.3). It doesn’t make any sense to count by bins of 8.3 points, so we may use 8, or 10, or 12. Next we look at where to start. The first number is 8. It doesn’t make any sense to start counting at 8 either, or to end at 91. We will probably want to start from 0 and end at 100, counting by 10’s should work nicely.

Mark horizontal axis: Mark your scale along the horizontal axis to cover your entire range and to count by your decided upon bin width. Include numbers.

Count number of values within each bin: How many values falls between 0 and <10? One, so we make the bin one unit tall. Between 10 and <20? Two, so we make the bin two units tall, etc. A frequency table may be helpful here. You need to know how tall to make each bin. You especially need to know how tall to make the tallest of the bins.

Mark vertical axis: Your vertical axis needs to reach the height of the tallest bin. Mark your vertical axis by consistent steps so that it will reach the number needed. Include numbers.

Make your histogram: Make the bins the correct heights, shade or color them in, add labels including any units, a title, and a key if needed.



Note that the bins in this example are [0 to 10); [10 to 20); etc. This means that zero up to, but not including, 10 are in the first bin (9.999 would be in bin #1, but 10 would be in bin #2).

Box Plots

A box plot is another type of graph used to display data. A **box plot** divides a set of numerical data into quarters. It shows how the data are dispersed around a median, but does not show specific values in the data. It does not show a distribution in as much detail as does a histogram, but it clearly shows where the data is located. This type of graph is often used when the number of data values is large or when two or more data sets are being compared. The center and spread of the distribution are very obvious from the graph. It is easy to see the range of the values as well as how these values are distributed around the middle value. The smaller the box, the more consistent the data values are with the median of the data. The shape of the box plot will give you a general idea of the shape of the distribution, but a histogram plot will do this more accurately. Any outliers will show up as long whiskers. The box in the box plot contains the middle 50% of the data, and each ’whisker’ contains 25% of the data.

1. Determine the five-number summary (will be defined later)
2. Draw a horizontal (or vertical) axis on which the numbers obtained in step 1 can be located. Above this axis, mark the quartiles and the minimum and maximum with vertical (horizontal) lines.
3. Connect the quartiles to each other to make a box, and then connect the box to the minimum and maximum with lines.

The boxplot is useful to assess the symmetry of the data:

- If the data are fairly symmetric, the median line will be roughly in the middle of the IQR box and the whiskers will be similar in length.
- If the data are skewed, the median may not fall in the middle of the IQR box, and one whisker will likely be noticeably longer than the other.

The Five Number Summary

Minimum, maximum and quartiles together provide information on center and variation of the variable in a nice compact way. Written in increasing order, they comprise what is called the five-number summary of the variable. In order to construct the Box Plot, those numbers should be firstly specified as follow:

- The left edge of the box represents the first quartile Q1, while the right edge represents the third quartile Q3. Thus the box portion of the plot represents the interquartile range IQR, or the middle 50% of the observations.
- The line drawn through the box represents the median of the data.
- The lines extending from the box are called whiskers. The whiskers extend outward to indicate the lowest and highest values in the data set (excluding outliers).
- Extreme values, or outliers, are represented by dots. A value is considered an outlier if it is outside of the box (greater than Q3 or less than Q1) by more than 1.5 times the IQR.

Skewness

The following relations indicate skewness or symmetry:

- Q3 − Q2 > Q2 − Q1 and Mean > Median indicate positive skew
- Q3 − Q2 < Q2 − Q1 and Mean < Median indicate negative skew
- Q3 − Q2 = Q2 − Q1 and Mean = Median indicate symmetry

The measure of skewness is based on the third sample moment about the mean

$$m_3 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

This is affected by the units we measure x in. Hence, a dimensionless form is used, called the coefficient of skewness:

$$\text{coeff. of skew} = \frac{m_3}{s^3}$$

Introduction of Probability

The world around us is full of phenomena we perceive as random or unpredictable. We aim to model these phenomena as *outcomes* of some experiment, where you should think of *experiment* in a very general sense. The outcomes are elements of a *sample space* S, and subsets of S are called *events*. The events will be assigned a *probability*, a number between 0 and 1 that expresses how likely the event is to occur.

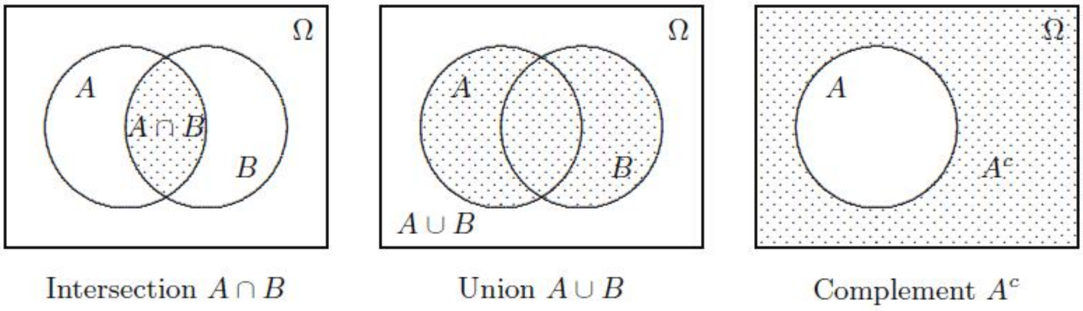
A **sample space** is a list of all the possible outcomes that may occur. What might happen when you flip a coin? You will either get heads or tails. What will happen when you roll a single die? You will either get a 1, 2, 3, 4, 5, or 6. The sample space for flipping a coin is S={heads, tails}. The sample space for rolling a die is S={1,2,3,4,5,6} On a coin flip, there are two **outcomes**, heads and tails. There are six different outcomes when considering the **event** of rolling a single die.

Subsets of the sample space are called *events*. We say that an event *A* occurs if the outcome of the experiment is an element of the set *A*. For example, in the birthday experiment we can ask for the outcomes that correspond to a long month, i.e., a month with 31 days. This is the event

L = *{*Jan, Mar, May, Jul, Aug, Oct, Dec*}*.

Events may be combined according to the usual set operations.

For example if *R* is the event that corresponds to the months that have the letter r in their (full) name (so *R* = {Jan, Feb, Mar, Apr, Sep, Oct, Nov, Dec}), then the long months that contain the letter r are $L \cap R$ = {Jan, Mar, Oct, Dec}. The set $L \cap R$ is called the *intersection* of *L* and *R* and occurs if both *L* and *R* occur. Similarly, we have the *union* $A \cup B$ of two sets *A* and *B*, which occurs if at least one of the events *A* and *B* occurs. Another common operation is taking complements. The event $Ac = \{\omega \in \Omega : \omega / \in A\}$ is called the *complement* of *A*; it occurs if and only if *A* does *not* occur. The complement of Ω is denoted ϕ, the empty set, which represents the impossible event. Figure 2.1 illustrates these three set operations. This figure is called Venn Diagram.



Intersection $A \cap B$ Union $A \cup B$ Complement A^c

Fig. 2.1 Venn Diagrams of intersection, union, and complement

Counting Techniques

The Fundamental Counting Principle states that if you wish to find the number of outcomes for a given situation, simply multiply the number of outcomes for each individual event. There are three basic counting techniques. They are multiplication rule, permutation and combination.

Multiplication Rule

If E1 is an experiment with n1 outcomes and E2 is an experiment with n2 possible outcomes, then the experiment which consists of performing E1 first and then E2 consists of n1n2 possible outcomes.

Permutation

A **permutation** is a specific order or arrangement of a set of objects or items. Consider a set of 4 objects. Suppose we want to fill 3 positions with objects selected from the above 4. Then the number of possible ordered arrangements is 24 and they are:

a b c	b a c	c a b	d a b
a b d	b a d	c a d	d a c
a c b	b c a	c b a	d b c
a c d	b c d	c b d	d b a
a d c	b d a	c d b	d c a
a d b	b d c	c d a	d c b

The number of possible ordered arrangements can be computed as follows:

Since there are 3 positions and 4 objects, the first position can be filled in 4 different ways. Once the first position is filled the remaining 2 positions can be filled from the remaining 3 objects. Thus, the second position can be filled in 3 ways. The third position can be filled in 2 ways. Then the total number of ways 3 positions can be filled out of 4 objects is given by

(4) (3) (2) = 24.

In general, if 'r' positions are to be filled from 'n' objects, then the total number of possible ways they can be filled are given by

$$\begin{aligned} & n(n-1)(n-2)\cdots(n-r+1) \\ & = \frac{n!}{(n-r)!} \\ & = {}_nP_r. \end{aligned}$$

Thus, nPr or P(n, r) represents the number of ways 'r' positions can be filled from 'n' objects. Each of the nPr arrangements is called a permutation of 'n' objects taken 'r' at a time.

General Formula for Permutation

1. The number of permutations of n different objects taken all at a time is n!.
2. The total number of arrangements of n different objects around a circle is (n-1)!.
3. The number of arrangements of n objects such that r_1 of them are of one kind, r_2 of them of a second kind, ….., r_k of the kth kind, is denoted by nPr_1, r_2, …, r_k and is given by:

$$nP_{r_1,r_2,\ldots\ldots,r_k} = \frac{n!}{r_1!\,r_2!\ldots\ldots r_k!} \quad , \; r_1 + r_2 + \ \ldots\ r_k = n$$

Combination

In permutation, order is important. But in many problems the order of selection is not important and interest centers only on the set of 'r' objects. Let 'c' denote the number of subsets of size 'r' that can be selected from n different objects. The 'r' objects in each set can be ordered in rPr ways. Thus we have

$${}_nP_r = c\ ({}_rP_r)$$

From this, we get

$$c = \frac{{}_nP_r}{{}_rP_r} = \frac{n!}{(n-r)!\,r!}$$

The number c is denoted by nCr or C(n, r). Thus, the above can be written as

$$ {}_{n}C_{r} = \frac{n!}{r!(n-r)!} $$

Each of the nCr unordered subsets is called a combination of 'n' objects taken 'r' at a time.

Binomial Expansion

It is customary to refer to the symbol C(n, r) as a binomial coefficient. To state and prove some theorems related to binomial coefficient, let us make the definition that C(n, r) = 0 whenever n is a positive integer and r is a positive integer greater than n.
For any positive integer n and r = 0, 1, 2, …, n, it is evident that:
C(n, r) = C(n, n-r)

Theorem 1

C(n, r) = C(n-1, r) + C(n-1, r-1) for r = 1, 2, ……, n-1.

Proof:

$$ C(n-1,r) + C(n-1,r-1) = \frac{(n-1)!}{r!\,(n-1-r)!} + \frac{(n-1)!}{(r-1)!\,(n-r)!} $$

$$ = \frac{(n-1)!}{(r-1)!\,(n-r-1)!}\left[\frac{1}{r} + \frac{1}{n-r}\right] $$

$$ = \frac{n(n-1)!}{r(r-1)!\,(n-r-1)!\,(n-r)} = \frac{n!}{r!\,(n-r)!} = C(n,r) $$

Theorem 2

$$ (a+b)^n = \sum_{r=0}^{n} C(n,r).\,a^r b^{n-r} $$

Proof

We shall prove the theorem by induction. The theorem is true for n=1 and n=2. Assume it is true for n=k, i.e.

$$ (a+b)^k = \sum_{r=0}^{k} C(k,r).\,a^r b^{k-r} $$

And we shall show that it is true for n=k+1. Thus

$$(a+b)^{k+1} = (a+b)(a+b)^k = (a+b)\left[\sum_{r=0}^{k} C(k,r).a^r b^{k-r}\right]$$

$$= (a+b)[C(k,0)a^0b^k + C(k,1)a^1b^{k-1} + \ldots + C(k,k)a^kb^0]$$

$$= C(k,0)a^0b^{k+1} + [C(k,0) + C(k,1)]a^1b^k$$

$$+ [C(k,1) + C(k,2)]a^2b^{k-1} + \ldots + C(k,k)a^{k+1}b^0$$

$$= b^{k+1} + C(k+1,1)a^1b^k + C(k+1,2)a^2b^{k-1} + \ldots + a^{k+1}$$

By theorem 1. So the theorem holds for all n.

Multinomial Expansion

Assume that there are n=7 students and that we wish to form 3 groups; 2 in the first, 3 in the second and 2 in the third group. Let n_1=2, n_2= 3 and n_3=2 indicate the numbers in the groups. Then

$n_1 + n_2 + n_3$ = n =7

There are C(n, n_i) = C(7, 2) = 21 different ways of selecting the first group of 2 students. After that there are C(n-n_i, n_2) = C(5, 3) = 10 different ways of selecting the second group of 3 students. Finally, there are C(n-n_1-n_2, n_3) = C(2, 2) = 1 way of selecting the remaining group.

By using the fundamental principle of counting, the total number of ways of selecting these three groups will be

$$C(n,n_1).C(n-n_1,n_2).C(n-n_1-n_2,n_3)$$

$$= \frac{n!}{n_1!\,(n-n_1)!}.\frac{(n-n_1)!}{n_2!\,(n-n_1-n_2)!}.\frac{(n-n_1-n_2)!}{n_3!\,0!} = \frac{n!}{n_1!\,n_2!\,n_3!}$$

$$= \frac{7!}{2!\,3!\,2!} = 210$$

In general, if we have n elements, and let n_1, n_2, …, n_k be positive integers with n_1+n_2+…+n_k=n, then there exist

$$\frac{n!}{n_1!\,n_2\ldots..!\,n_k!}$$

different ordered partitions of n objects into these k groups, and are denoted by

$$nC_{n_1,n_2,\ldots\ldots,n_k}\ or\ C(n;\ n_1,n_2,\ldots\ldots,n_k)$$

and is written as

$$C(n;\ n_1,n_2,\ldots\ldots,n_k) = \frac{n!}{n_1!\,n_2\ldots..!\,n_k!}$$

The equation above is referred to as a multinomial coefficient.

Theorem

$$(a_1 + a_2 + \cdots. + a_k)^n = \sum C(n;\ n_1,n_2,\ldots\ldots,n_k)a_1^{n_1}a_2^{n_2}\ldots\ldots.a_k^{n_k}$$

The summation is overall possible sets of integers $(n_1,n_2,\ldots..,n_k)$ such that

$0 \leq n_i \leq n, \quad i = 1,2,\ldots..,n$

and $n_1 + n_2 + \ldots..+ n_k = n$

there are C(n+k-1, n) terms in the expansion of equation above.

Probability

One of the fundamental tools of statistics is Probability, which had its formal beginning with games of chance in the seventeenth century. Probability is the study of random or nondeterministic experiments (that is, we wish to consider some experiment results and the repetition of the experiment does not always produce the same results). For example, when a coin is flipped, the chance that it comes up heads is 50%. Probabilities can be expressed as decimals, fractions, percents, or ratios. We could have said the probability of flipping heads is , 0.5, 1/2 , 50% or 1:2. Each of these conveys the idea that we should expect to get a heads half of the time. Probabilities only give us an idea of what to expect in the long run.

Now suppose we flip a coin 10 times in a row and get heads each time. The next coin flip is still a **random event** because while we cannot tell for certain what the next flip will be, we can be certain that about 50% of all tosses over a long set of tosses will be heads. Some people think that we are on a roll so we are more likely to get another heads. Others will say that getting tails is more likely because we are due to get tails.

The truth is that we cannot tell what will happen on the next flip. The only thing we know for certain is that there is a 50% chance that the coin will be heads on its next flip. If we continue to flip this same coin hundreds of times, we would expect the percent of heads to get closer and closer to 50%.

The **Law of Large Numbers** tells us that despite the results on a small number of flips, we will eventually get closer to the **theoretical probability**. The outcomes in any random event will always get close to the theoretical probability if the event is repeated a large number of times. When calculating a probability, we divide the number of favorable outcomes (outcomes we are interested in) by the total number of outcomes. In other words, the probability that outcome 'A' occurs is found by the formula:

P(A) = No. of favorable outcomes / total No. o f outcomes .

Consider a standard deck of 52 playing cards. If we asked the question "What is the probability of being dealt a face card (jack, queen, or king)?", we would need to count how many cards are face cards and then divide by the total number of cards, 52. In this situation there are 12 face cards and 52 cards overall so our probability of getting a face card is 12/52 = 3/13 = 0.23. The list below are some rules in probability:

1. The probability of a sure thing is 1.
2. The probability of an impossible outcome is 0.
3. The sum of the probabilities of all possible outcomes is 1.
4. The probability for any random event must be somewhere from 0 to 1.

We notate the probability of event 'A' happening as P(A). For example, the probability of rolling a three on a six-sided die can be written $P(3) = 1/6$. Sometimes we are interested in the probability of an event not occurring. This is called the **complement** of the event. We can write the probability of the complement of event 'A' happening as P(˜A), P(not A), or $P(A^c)$. The formula for the complement of an event is $P(not\ A) = 1 - P(A)$. On our die rolling question, P(˜3)= 1 - $P(3)$ = 1 – 1/6 = 5/6 . In other words, there is a 5/6 chance of the dice not landing on a 3. It is important to notice that the probability of an event happening and the probability of its complement always add up to 1.

Kinds of Probability

There are three basic ways of classifying probability. These three represent rather different conceptual approaches to the study of probability theory, and in fact experts disagree about which approach is the proper one to use. The three kinds are:

1. The classical approach
2. The relative frequency approach
3. The subjective approach

The Classical Approach

If a random experiment, whose sample space 'S', can result in 'n' mutually exclusive and equally likely outcomes (i.e. each outcome has the same chance for occurrence) and if n(A) of these outcomes have an attribute 'A', then the probability of 'A', denoted by P(A), is defined by,

$$P(A) = \frac{n(A)}{n} = \frac{number\ of\ outcomes\ favourable\ to\ A}{total\ number\ of\ possible\ outcomes}$$

Classical probability is often called a prior probability because, if we keep using examples like tossing a fair coin, tossing a die and drawing a card from ordinary deck of cards, we can state the answer in advance without conducting a large number of trials.

Note

If A and B are two mutually exclusive events, then

P(A or B) = P(A∪B) = P(A) + P(B)

Relative Frequency Approach

We may not be able in the classical approach to answer questions like the following:

1. What is the probability that a man would live more than 70 years?
2. What is the probability that student chosen at random from a class of 100 students will pass the exam?

Without experimentation, we may not be able to do this in advance. So another kind of approach may be more useful. It is the relative frequency or posterior probability, which defines the probability as the proportion of times that an event occurs in the long run when the conditions of performing the experiment under it are stable.

If the number of independent trials is infinite, then we define the probability of an event E as

$$P(E) = \lim_{n\to\infty} \frac{h}{n}$$

Where 'h' and 'n' denote the number of occurrences of 'E' and the number of independent trials, respectively. The ratio h/n is called the relative frequency.

Subjective Probability

Subjective probability is based on the personal beliefs of the person making the probability assessment. In fact, subjective probability can be defined as the probability assigned to an event by an individual, based on whether evidence is available. This evidence may be in the form of relative frequencies of past occurrences, or it may be just an educated guess.

Compound and Independent Events

From the previous section, we found that it is quite straightforward to calculate probabilities for simple situations. What happens when we calculate probabilities from multiple events? For example, suppose you roll a single die and then flip a coin. What are the chances that the die comes up with a 5 and the coin gives you a heads? A situation that asks you to calculate probabilities for a situation that involves two or more events or steps is called a **compound event.** We will try to find out how to handle these types of situations by examining several situations and then making a conclusion.

We could multiply the probabilities of each individual event to get the probability of both events happening. This is always true of independent events. In other words, suppose we want the probability of both some outcome ’A’ from one event and some outcome ’B’ from a second event that is independent of the first event. If the probability of our first outcome is P(A) and the probability of our second outcome is P(B), then the probability of both A and B happening is P(A and B) = *P*(*A*) x *P*(*B*).

For Independent Events

$$P(A \,\&\, B) = P(A) \cdot P(B)$$

Note

Sometimes there are situations in which two different outcomes cannot occur at the same time. For example, if you roll a single die one time and you wish to find the probability of getting an even number and a 3 on that one roll. These two outcomes cannot occur at the same time. When it is impossible for two outcomes to occur at the same time, we say the outcomes are **mutually exclusive** or **disjoint.** If outcomes ’A’ and ’B’ are mutually exclusive then it is impossible for outcome ’A’ and ’B’ to happen at the same time, or P(A and B)=0. However, if ’A’ and ’B’ are mutually exclusive then P(A or B)=P(A)+P(B). Using proper notation we have $P(A \cup B) = P(A) + P(B)$. *Remember, this is only true if the two outcomes are mutually exclusive.*

For Mutually Exclusive Events

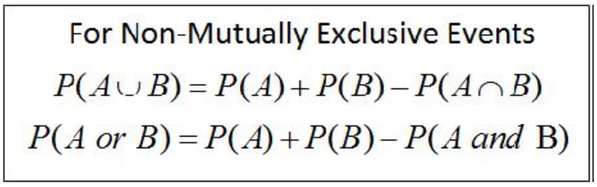
$$P(A \cup B) = P(A) + P(B)$$

$$P(A \text{ or } B) - P(A) + P(B)$$

For any sequence of mutually exclusive events A_1, A_2, ….. of Sample Space, that is, events for which A_iA_j = ϕ for i≠j,

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

For Non-mutually exclusive events, the formula for the union of two outcomes will be equal to:



For Non-Mutually Exclusive Events

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } \mathrm{B})$$

Tree Diagrams and Probability Models

As we advance through probability, it becomes very apparent that we need to be quite organized with our problems as they become more complex. In this section we will use tree diagrams to help us calculate probabilities for given situations. **Tree diagrams** are a visual aid that can help us break down a situation and calculate probabilities. There are two key principles that we must observe for all tree diagrams. First of all, to find the total probability for any given branch on a tree, multiply the individual probabilities along that branch. Secondly, the sum of the probabilities from the ends of each branch must total to 1. We will examine several examples of probabilities using tree diagrams in order to solidify our understanding of this concept.

Conditional Probabilities

It is quite easy to calculate simple probabilities. What is the chance of rolling a 4 with a single die? What is the chance of being dealt a queen from a deck of cards? We are now going to focus on conditional probabilities. A **conditional probability** is a probability in which a certain prerequisite condition has already been met.

If we let *'E'* and *'F'* denote, respectively, the event that the sum of numbers on both dice is 8 and the event that the first die is a 3, then the probability just obtained is called the *conditional probability that 'E' occurs given that 'F' has occurred* and is denoted by

$$P(E|F)$$

A general formula for *P(E|F)* that is valid for all events *'E'* and *'F'* is derived in the same manner: If the event '*F'* occurs, then, in order for *'E'* to occur, it is necessary that the actual occurrence be a point both in *'E'* and in *'F'*; that is, it must be in *'EF'*. Now, since we know that *'F'* has occurred, it follows that *'F'* becomes our new, or reduced, sample space; hence, the probability that the event *EF* occurs will equal the probability of *EF* relative to the probability of *'F'*. That is, we have the following definition.

If *P(F)* > 0, then

$$P(E|F) = P(EF)/P(F)$$

Note

1. We can define the conditional probability for an event 'E' given that event F has occurred by the formula

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

provided that P(F)>0

2. Another way we can look at conditional probabilities is through the use of **two-way tables** or **contingency tables**. These are often referred to as two-way tables because there are two distinct pieces of information gathered in these tables. For example, we may record how many siblings you have and in how many activities you participate in school. Two-way tables can be filled in either using counts or probabilities.

Theorem of Total Probability

Suppose that events B_1, B_2, . . . , and B_n are mutually exclusive and exhaustive (i.e. *S* = B_1 + B_2 + ….. + B_n). Then, for an arbitrary event *A*

$$\boxed{\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n) \\ &= \sum_{j=1}^{n} P(A|B_j)P(B_j). \end{aligned}}$$

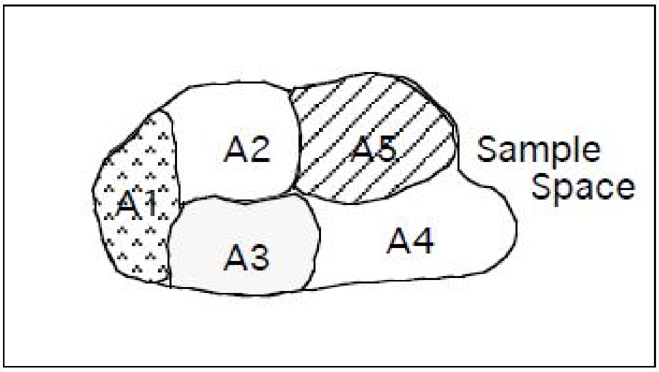
Baye's Theorem

There are many situations where the ultimate outcome of an experiment depends on what happens in various intermediate stages. This issue is resolved by the Bayes' Theorem.

Let S be a set and let $P = \{A_i\}_{i=1}^{m}$ be a collection of subsets of S. The collection P is called a partition of S if

$$(a) \quad S = \bigcup_{i=1}^{m} A_i$$

$$(b) \quad A_i \cap A_j = \emptyset \quad \text{for } i \neq j.$$



If the events $\{B_i\}_{i=1}^{m}$ constitute a partition of the sample space S and P(B_i) ≠ 0 for i = 1, 2, ...,m, then for any event A in S

$$P(A) = \sum_{i=1}^{m} P(B_i)\, P(A/B_i).$$

Proof

Let S be a sample space and A be an event in S. Let $\{B_i\}_{i=1}^{m}$ be any partition of S. Then

$$A = \bigcup_{i=1}^{m} (A \cap B_i)$$

Thus

$$P(A) = \sum_{i=1}^{m} P(A \cap B_i)$$
$$= \sum_{i=1}^{m} P(B_i)\, P(A/B_i)$$

Now, If the events $\{B_i\}_{i=1}^{m}$ constitute a partition of the sample space S and P(B_i) ≠ 0 for i = 1, 2, ...,m, then for any event A in S such that P(A) ≠ 0

$$P(B_k/A) = \frac{P(B_k)\, P(A/B_k)}{\sum_{i=1}^{m} P(B_i)\, P(A/B_i)} \quad k = 1, 2, ..., m$$

This Theorem is called Bayes Theorem. The probability P(B_k) is called prior probability. The probability P(B_k/A) is called posterior probability.

Random Variables

In many random experiments, the elements of sample space are not necessarily numbers. For example, in a coin tossing experiment the sample space consists of
S = {Head, Tail}.

Statistical methods involve primarily numerical data. Hence, one has to 'mathematize' the outcomes of the sample space. This mathematization, or quantification, is achieved through the notion of random variables.

Consider a random experiment whose sample space is S. A random variable X is a function which assigns to each element s ϵ S a real number X(s) in the set of real numbers R.

In a particular experiment a random variable X would be some function that assigns a real number X(s) to each possible outcome s in the sample space. Given a random experiment, there can be many random variables. This is due to the fact that given two (finite) sets A and B, the number of distinct functions one can come up with is $|B|^{|A|}$. Here |A| means the cardinality of the set A.

Random variable is not a variable. Also, it is not random. Thus someone named it inappropriately. A random variable is neither random nor variable, it is simply a function. The values it takes on are both random and variable.

The space of the random variable X will be denoted by R_x and represents the set {x ϵ R| x = X(s), s ϵ S}.

Notes:

- Now, we introduce some notations. By (X = x) we mean the event {s ϵ S |X(s) = x}. Similarly, (a < X < b) means the event {s ϵ S | a < X < b} of the sample space S.

- There are three types of random variables: discrete, continuous, and mixed. However, in most applications we encounter either discrete or continuous random variable. The random variable X is called a discrete random variable if it is defined over a sample space having a finite or a countably infinite number of sample points. In this case, random variable X takes on discrete values, and it is possible to enumerate all the values it may assume. In the case of a sample space having an uncountably infinite number of sample points, the associated random variable is called a continuous random variable, with its values distributed over one or more continuous intervals on the real line. In this chapter we only treat these two types of random variables. First, we consider the discrete case and then we examine the continuous case.

Probability Distributions

The behavior of a random variable is characterized by its probability distribution, that is, by the way probabilities are distributed over the values it assumes. A probability distribution function and a probability mass function are two ways to characterize this distribution for a discrete random variable. They are equivalent in the sense that the knowledge of either one completely specifies the random variable. The corresponding functions for a continuous random variable are the probability distribution function, defined in the same way as in the case of a discrete random variable, and the probability density function. The definitions of these functions now follow.

Probability Distribution Function (PDF)

Given a random experiment with its associated random variable X and given a real number 'x', let us consider the probability of the event {s: X(s) ≤ x}or simply P(X ≤ x), This probability is clearly dependent on the assigned value x. The function

$$F_x(x) = P(X \leq x)$$

is defined as Probability Distribution Function (PDF) or simply Distribution Function of X. the subscript X in the equation above identifies the random variable. The PDF is thus the probability that X will assume a value lying in a subset of S, the subset being point 'x' and all points lying to the ‘left’ of 'x'. As 'x' increases, the subset covers more of the real line, and the value of PDF increases until it reaches 1. The PDF of a random variable thus accumulates probability as increases, and the name Cumulative Distribution Function(CDF) is also used for this function.

In view of the definition and the discussion above, we give below some of the important properties possessed by a PDF.

- It exists for discrete and continuous random variables and has values between 0 and 1.

- It is a nonnegative, continuous-to-the-left, and nondecreasing function of the real variable x. Moreover, we have

$$F_X(-\infty) = 0 \qquad and \qquad F_x(+\infty) = 1$$

- If 'a' and 'b' are two real numbers such that a < b, then

$$P(a < X \leq b) = F_X(b) - F_x(a)$$

  This relation is a direct result of the identity

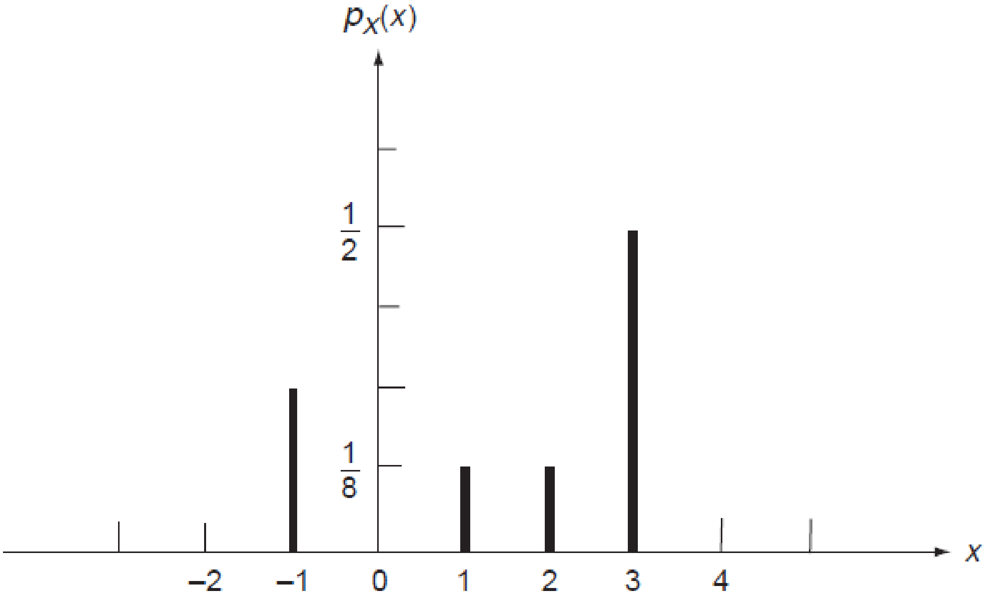
  $P(X \leq b) = P(X \leq a) + P(a < X \leq b)$

- $P(X > b) = 1 - F_x(b)$

PROBABILITY MASS FUNCTION FOR DISCRETE RANDOM VARIABLES

Let be a discrete random variable that assumes at most a countably infinite number of values x_1, x_2 . . . with nonzero probabilities. If we denote $P(X = x_i) = p(x_i)$, $i = 1, 2$ … .. , then, clearly,

$$\left.\begin{array}{l} 0 < p(x_i) \leq 1, \text{ for all i;} \\ \sum_i p(x_i) = 1. \end{array}\right\}$$



Probability mass function of $X, p_X(x)$ for the random variable defined in Example

Note: the function

$$\boxed{p_X(x) = P(X = x).}$$

is defined as the probability mass function (pmf) of . Again, the subscript X is used to identify the associated random variable.

We can observe that, like $F_X(x)$, the specification of $p_X(x)$ completely characterizes random variable ; furthermore, these two functions are simply related by:

$$\boxed{\begin{aligned} p_X(x_i) &= F_X(x_i) - F_X(x_{i-1}), \\ F_X(x) &= \sum_{i=1}^{i:x_i \leq x} p_X(x_i), \end{aligned}}$$

The upper limit for the sum in the above Equation means that the sum is taken over all satisfying $x_i \leq x$. Hence, we see that the PDF and pmf of a discrete random variable contain the same information; each one is recoverable from the other.

PROBABILITY DENSITY FUNCTION FOR CONTINUOUS RANDOM VARIABLES

For a continuous random variable , its PDF, $F_x(x)$ is a continuous function of x and the derivative

$$\boxed{f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}}$$

exists for all x. the function $f_x(x)$ is called the probability density function (pdf), or simply the density function of X.

Since $F_x(x)$ is monotone non-decreasing, we clearly have

$$f_X(x) \geq 0 \quad \text{for all } x.$$

Note: Additional properties of $f_x(x)$ are mentioned below,

$$\boxed{F_X(x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u,}$$

$$\left.\begin{aligned} &\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1, \\ &P(a < X \le b) = F_X(b) - F_X(a) = \int_a^b f_X(x)\,\mathrm{d}x. \end{aligned}\right\}$$

Some Special Discrete Distributions

Discrete random variables are often classified according to their probability mass functions. In this section, we explore some frequently encountered discrete distributions and study their important characteristics.

Bernoulli Distribution

A Bernoulli trial is a random experiment in which there are precisely two possible outcomes, which we conveniently call ‘failure’ (F) and ‘success’ (S). We can define a random variable from the sample space {S, F} into the set of real numbers as follows:

X(F) = 0 X(S) = 1

The probability mass function of Bernoulli random variable is
f(0) = P(X=0) = 1- p
f(1) = P(X=1) = p
where p denotes the probability of success. Hence

$$f(x) = p^x\,(1-p)^{1-x}, \qquad x = 0, 1.$$

We denote this distribution by Ber(p).

Binomial Distribution

Consider a fixed number n of mutually independent Bernoulli trails. Suppose these trials have same probability of success, say p. A random variable X is called a binomial random variable if it represents the total number of successes in n independent Bernoulli trials.

Now we determine the probability mass function of a binomial random variable. Recall that the probability mass function of X is defined as

$$f(x) = P(X = x).$$

Thus, to find the probability mass function of X we have to find the probability of x successes in n independent trails.

If we have x successes in n trails, then the probability of each n-tuple with x successes and n − x failures is

$$p^x (1 - p)^{n-x}.$$

However, there are nCx tuples with x successes and n − x failures in n trials. Hence

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}.$$

Therefore, the probability mass function of X is

$$f(x) = \binom{n}{x} p^x (1 - p)^{n-x}, \qquad x = 0, 1, ..., n.$$

We will denote a binomial random variable with parameters p and n as X = BIN(n, p).

Geometric Distribution

Another event of interest arising from Bernoulli trials is the number of trials to (and including) the first occurrence of success. If *X* is used to represent this number, it is a discrete random variable with possible integer values ranging from one to infinity. Its pmf is easily computed to be

$$\boxed{\begin{aligned} p_X(k) &= P(\underbrace{FF\ldots F}_{k-1} S) = P\underbrace{(F)P(F)\ldots P(F)}_{k-1} P(S) \\ &= q^{k-1}p, \quad k = 1, 2, \ldots. \end{aligned}}$$

Where 'p' is the probability of success and 'q' is the probability of failed. This distribution is known as the *geometric distribution* with parameter *p*, where the name stems from its similarity to the familiar terms in geometric progression.

The corresponding probability distribution function is

$$F_X(x) = \sum_{k=1}^{m \leq x} p_X(k) = p + qp + \cdots + q^{m-1}p$$
$$= (1-q)(1+q+q^2+\cdots+q^{m-1}) = 1 - q^m;$$

where *m* is the largest integer less than or equal to *x*.

Some Special Continuous Distributions

Let us turn our attention to some important continuous probability distributions. Physical quantities such as time, length, area, temperature, pressure, load, intensity, etc., when they need to be described probabilistically, are modeled by continuous random variables. A number of important continuous distributions are introduced in this section.

Uniform Distribution

Let the random variable X denote the outcome when a point is selected at random from an interval [a, b]. We want to find the probability of the event X ≤ x, that is we would like to determine the probability that the point selected from [a, b] would be less than or equal to x. Hence,

$$P(X \leq x) = \frac{\text{length of [a, x]}}{\text{length of [a, b]}}.$$

Thus, the cumulative distribution function F is

$$F(x) = P(X \leq x) = \frac{x-a}{b-a}, \qquad a \leq x \leq b,$$

where a and b are any two real constants with a < b. To determine the probability density function from cumulative density function, we calculate the derivative of F(x). Hence

$$f(x) = \frac{d}{dx}F(x) = \frac{1}{b-a}, \qquad a \leq x \leq b.$$

Note:

A random variable X is said to be uniform on the interval [a, b] if its probability density function is of the form

$$f(x) = \frac{1}{b-a}, \qquad a \leq x \leq b,$$

where a and b are constants. We denote a random variable X with the uniform distribution on the interval [a, b] as X ~ UNIF or U(a, b).

Exponential Random Variable

A continuous random variable whose probability density function is given, for some $\lambda > 0$, by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

is said to be an exponential random variable (or, more simply, is said to be exponentially distributed) with parameter 'λ'. The cumulative distribution function F(a) of an exponential random variable is given by

$$\begin{aligned} F(a) &= P\{X \leq a\} \\ &= \int_0^a \lambda e^{-\lambda x}\,dx \\ &= -e^{-\lambda x}\Big|_0^a \\ &= 1 - e^{-\lambda a} \quad a \geq 0 \end{aligned}$$

Note that $F(\infty) = \int_0^\infty \lambda \mathrm{e}^{-\lambda \mathrm{x}} \mathrm{dx} = 1$ as, of course, it must.

Normal (or Gaussian) Distribution

A random variable X is said to have a normal distribution if its probability density function is given by

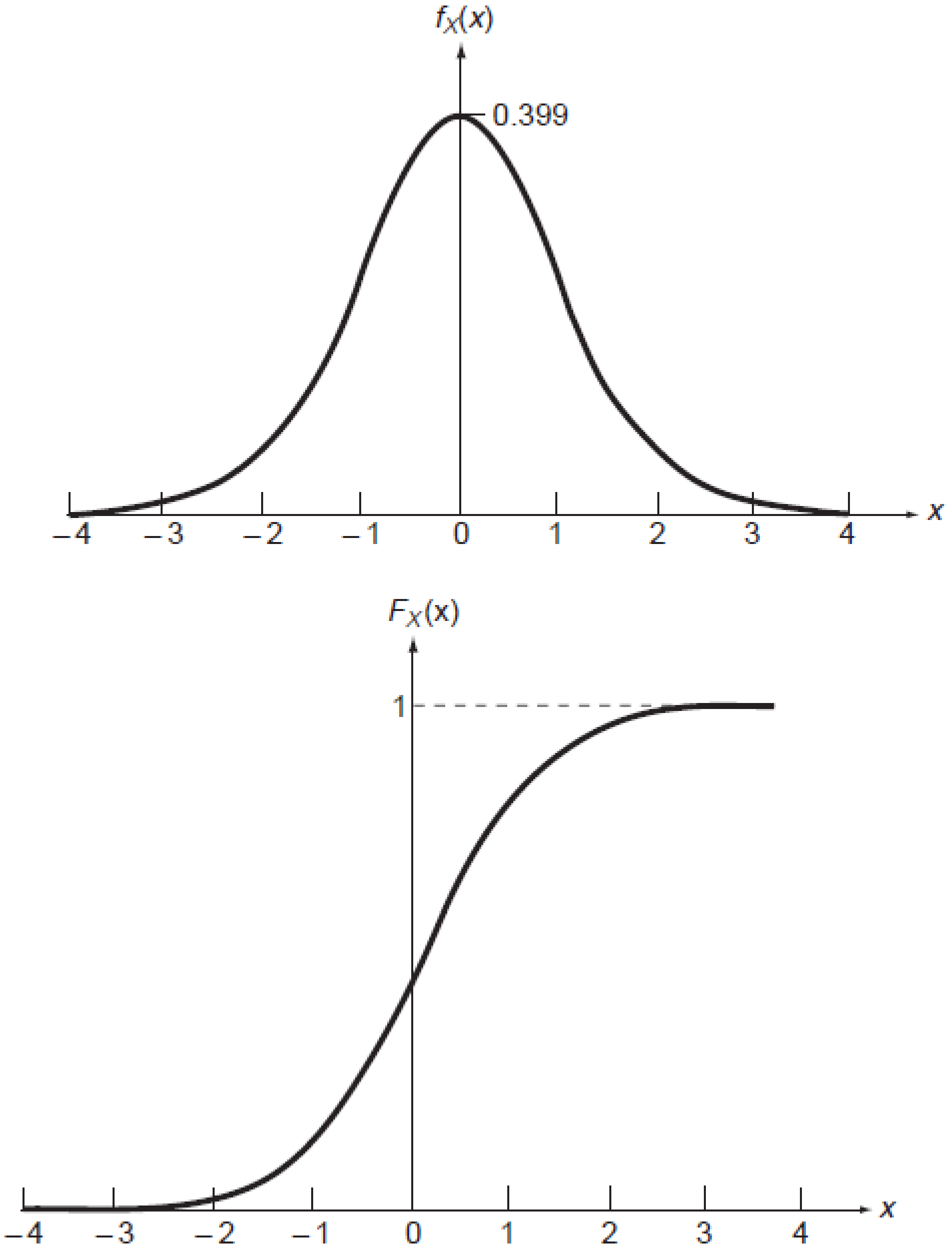
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < x < \infty,$$

where $-\infty < \mu < \infty$ and $0 < \sigma^2 < \infty$ are arbitrary parameters. If X has a normal distribution with parameters μ and σ^2, then we write X ~ N(μ, σ^2). The values μ and σ^2 represent the average value and the variance of X, respectively.

Its corresponding PDF is

$$\boxed{F_X(x) = \frac{1}{(2\pi)^{1/2}\sigma} \int_{-\infty}^{x} \exp\left[-\frac{(u-m)^2}{2\sigma^2}\right] du, \quad -\infty < x < \infty,}$$

which cannot be expressed in closed form analytically but can be numerically evaluated for any *x*. The pdf and PDF expressed by Equations above are graphed in Figures below for μ= 0 and σ =1. The graph of *f X* (*x*) in this particular case is the well-known bell-shaped curve, symmetrical about the origin. An important implication of the preceding result is that if X is normally distributed with parameters μ and $\sigma2$, then $Z = (X - \mu)/\sigma$ is normally distributed with parameters 0 and 1. Such a random variable is said to be a *standard*, or a *unit*, normal random variable.



Probability density function $f_X(x)$ and probability distribution function, $F_X(x)$, of X for $\mu = 0$ and $\sigma = 1$

$N(\mu, \sigma^2)$ distributed random variable can be turned into an $N(0, 1)$ distributed random variable by a simple transformation. As a consequence, a table of the $N(0, 1)$

distribution suffices. The latter is called the *standard* normal distribution, and because of its special role the letter 'ϕ' has been reserved for its probability density function:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad \text{for } -\infty < x < \infty.$$

Note that ϕ is symmetric around zero: $\phi(-x) = \phi(x)$ for each x. The corresponding distribution function is denoted by Φ. The table for the standard normal distribution (see the Table) does not contain the values of $\Phi(a)$, but rather the so-called *right tail probabilities* $1-\Phi(a)$. If, for instance, we want to know the probability that a standard normal random variable Z is smaller than or equal to 1, we use that $P(Z \leq 1) = 1 - P(Z \geq 1)$. In the table we find that $P(Z \geq 1) = 1-\Phi(1)$ is equal to 0.1587. Hence $P(Z \leq 1) = 1-0.1587 = 0.8413$. With the table you can handle tail probabilities with numbers a given to two decimals. To find, for instance, $P(Z > 1.07)$, we stay in the same row in the table but move to the seventh column to find that $P(Z > 1.07) = 0.1423$.

The corresponding Distribution Function Φ(z) = P{Z≤z} is given by

$$\Phi(z) = \int_{-\infty}^{z} \phi(u)du = \int_{-\infty}^{z} \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}} du$$

Note:

If Z is a standard normal random variable, then

$$P\{Z \leq -x\} = P\{Z > x\} \qquad -\infty < x < \infty$$

Since $Z = (X - \mu)/\sigma$ is a standard normal random variable whenever X is normally distributed with parameters μ and σ^2, it follows that the distribution function of X can be expressed as

$$F_X(a) = P\{X \leq a\} = P\left(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Note:

P(Z< -a) = P(Z > a) = 1- P(Z ≤ a) = 1 – Φ(a)

Φ(-a) = 1 - Φ(a)

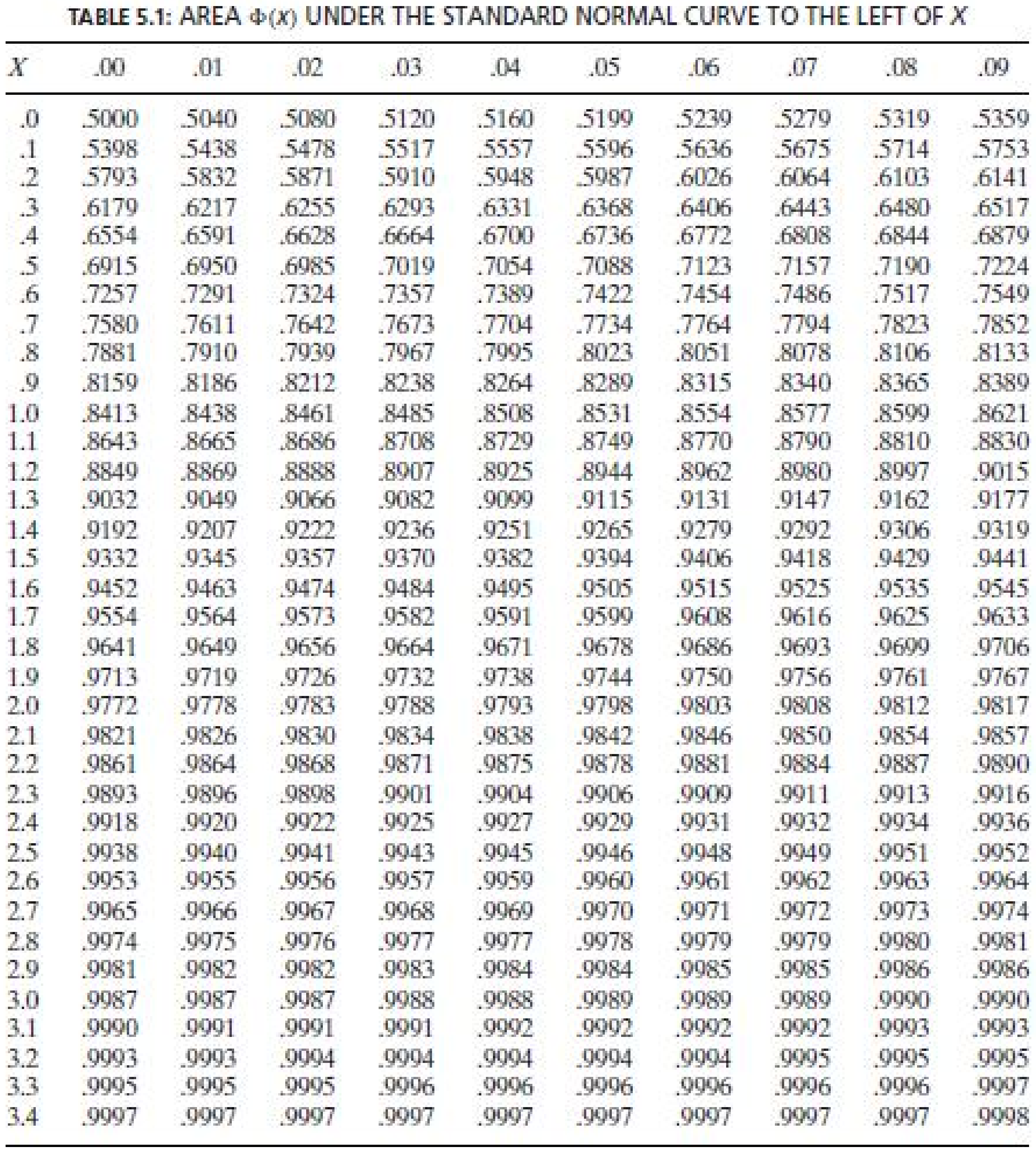


TABLE 5.1: AREA $\Phi(x)$ UNDER THE STANDARD NORMAL CURVE TO THE LEFT OF X

X	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Expectations

Random variables are complicated objects, containing a lot of information on the experiments that are modeled by them. If we want to summarize a random variable by a *single number*, then this number should undoubtedly be its *expected value*. The expected value, also called the *expectation* gives the center of the distribution of the random variable.

Expected Value for Discrete Random Variables

One of the most important concepts in probability theory is that of the expectation of a random variable. If '*X*' is a discrete random variable having a probability mass function '*p(x)*', then the *expectation*, or the *expected value*, of '*X*', denoted by '*E*[*X*]', is defined by

$$E[X] = \sum_{x:p(x)>0} xp(x)$$

In words, the expected value of '*X*' is a weighted average of the possible values that *X* can take on, each value being weighted by the probability that *X* assumes it. For instance, on the one hand, if the probability mass function of *X* is given by

$$p(0) = \frac{1}{2} = p(1)$$

then

$$E[X] = 0\left(\frac{1}{2}\right) + 1\left(\frac{1}{2}\right) = \frac{1}{2}$$

is just the ordinary average of the two possible values, 0 and 1, that *X* can assume. On the other hand, if

$$p(0) = \frac{1}{3} \qquad p(1) = \frac{2}{3}$$

then

$$E[X] = 0\left(\frac{1}{3}\right) + 1\left(\frac{2}{3}\right) = \frac{2}{3}$$

is a weighted average of the two possible values 0 and 1, where the value 1 is given twice as much weight as the value 0, since $p(1) = 2p(0)$. Now, consider a random variable *X* that must take on one of the values x_1, x_2, . . . x_n with respective probabilities $p(x_1)$, $p(x_2)$, . . . , $p(x_n)$, and think of *X* as representing our winnings in a single game of chance. That is, with probability $p(x_i)$ we shall win ' x_i 'units $i = 1, 2, . .$

. , *n*. By the frequency interpretation, if we play this game continually, then the proportion of time that we win 'x_i' will be $p(x_i)$. Since this is true for all 'i', $i = 1, 2, . . . , n$, it follows that our average winnings per game will be

$$\sum_{i=1}^{n} x_i p(x_i) = E[X]$$

Note

Let g(*X*) be a real-valued function of a random variable *X*. The *mathematical expectation*, or simply *expectation*, of g(*X*), denoted by E{g(X)} is defined by

$$\boxed{E\{g(X)\} = \sum_i g(x_i)p_X(x_i),}$$

if *X* is discrete, where $x_1, x_2, . . .$ are possible values assumed by *X*.

Expected Value for Continuous Random Variable

If *X* is a continuous random variable having probability density function *f (x)*, then, because

$$f(x)\,dx \approx P\{x \le X \le x + dx\} \quad \text{for } dx \text{ small}$$

it is easy to see that the analogous definition is to define the expected value of *X* by

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,dx$$

Notes

Let us note some basic properties associated with the expectation operator.

1. If 'a' and 'b' are constant, then

$$E[aX + b] = aE[X] + b$$

2. For any constant 'c' and any functions *g*(*X*) and *h*(*X*) for which expectations exist, we have

$$E\{c\} = c,$$
$$E\{cg(X)\} = cE\{g(X)\},$$
$$E\{g(X) + h(X)\} = E\{g(X)\} + E\{h(X)\},$$
$$E\{g(X)\} \leq E\{h(X)\}, \quad \text{if } g(X) \leq h(X) \text{ for all values of } X.$$

Moments of a Single Random Variable

Let $g(X) = X^n$, $n = 1, 2, \ldots$; the expectation $E\{X^n\}$, when it exists, is called the *n*th *moment* of *X*. It is denoted by 'α_n' and is given by

$$\alpha_n = E\{X^n\} = \sum_i x_i^n p_X(x_i), \quad \text{for } X \text{ discrete;}$$
$$\alpha_n = E\{X^n\} = \int_{-\infty}^{\infty} x^n f_X(x)\mathrm{d}x, \text{ for } X \text{ continuous.}$$

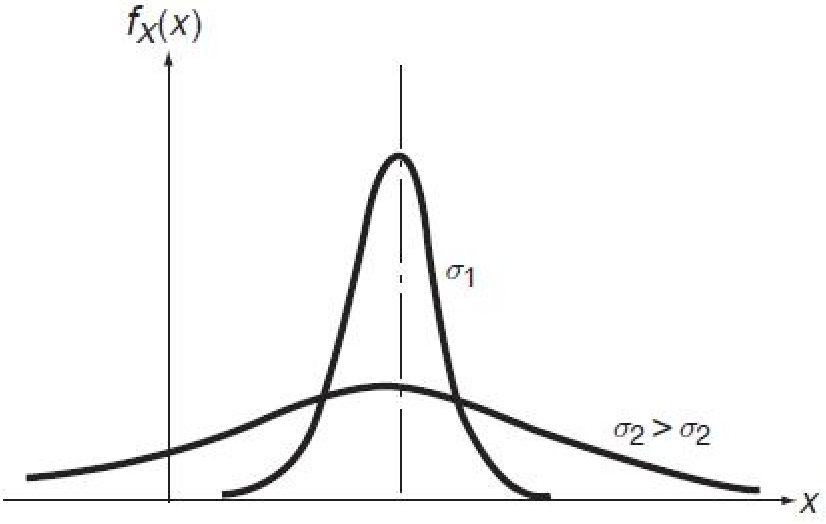
Note that the first moment is called mean or expected value.

Note

The *central moments* of random variable *X* are the moments of *X* with respect to its mean. Hence, the *n*th *central moment* of *'X'*, 'μ_n', is defined as

$$\mu_n = E\{(X - m)^n\} = \sum_i (x_i - m)^n p_X(x_i), \quad X \text{ discrete;}$$
$$\mu_n = E\{(X - m)^n\} = \int_{-\infty}^{\infty} (x - m)^n f_X(x)\mathrm{d}x, \quad X \text{ continuous.}$$

The *variance* of *X* is the second central moment, μ_2, commonly denoted by σ^2_x or simply σ^2 or var(*X*). It is the most common measure of dispersion of a distribution about its mean. Large values of σ^2_x imply a large spread in the distribution of *X* about its mean. Conversely, small values imply a sharp concentration of the mass of distribution in the neighborhood of the mean. This is illustrated in Figure below in which two density functions are shown with the same mean but different variances. When $\sigma^2_x = 0$, the whole mass of the distribution is concentrated at the mean. In this extreme case, $X = m_X$ with probability 1.



Density functions with different variances σ₁ and σ₂

We note two other properties of the variance of a random variable X which can be similarly verified. They are:

$$\left.\begin{aligned} \mathrm{var}(X+c) &= \mathrm{var}(X), \\ \mathrm{var}(cX) &= c^2\mathrm{var}(X), \end{aligned}\right\}$$

where c is any constant.

An important relation between the variance and simple moments is

$$\sigma^2 = \alpha_2 - m^2.$$

That is

$$\boxed{\mathrm{Var}(X) = E[X^2] \; - \; (E[X])^2}$$

Sampling

The purpose of sampling is to enable us to make inferences about a population after inspecting only a portion (a *sample*) of that population. Such factors as cost, time, destructive testing, and infinite populations make sampling preferable to making a complete inspection (census, complete enumeration) of a population. Usually, there are some numerical characteristics about the population which the investigator wants to know. Such numerical facts are called *parameters*; e.g., the population size, population mean, a proportion of some attribute, and variability in the population, etc.

A *Parameter* is a numerical value, or a characteristic, of a population. The parameters cannot be determined exactly, but can only be estimated from a sample by quantities called statistics. A Statistic is a numerical summary, or a characteristic, of a sample.

A very important and basic type of probability sampling is the *Simple Random Sampling*. Variations of simple random sampling include systematic, stratified, and cluster sampling. These sampling methods make the methods of sampling in statistics.

Simple random sampling: Simple random sampling selects samples, of size n, by methods that allow each possible sample of size n to have an equally likely chance or equal probability of being considered. In addition, each unit, in the entire population, has an equal chance of being included in the sample in each single drawing.

A Systematic Sample is obtained by selecting every kth individual from the population. The first individual selected correspond to a random number between 1 and k. Steps in Systematic sampling

1. Make sure that you have a finite population of size N, and enumerate each individual.

2. Decide on how large your sample will be, n.

3. Calculate N/n, and round, up or down, to an integer. Let that integer be k.

4. Randomly select a number between 1 and k, call that number j.

5. Thus the sample will consist of the following enumerated individuals:

j, j+k, j+2k, …j+(n-1)k.

A Stratified Sample is obtained by dividing the population into separate homogeneous categories, or groups, that do not overlap. These are called Strata (the singular is Stratum) and then take a sample by simple random sampling from each stratum. These chosen subsamples will form the stratified sample needed.

A Cluster Sample is obtained like the stratified sample by dividing the population into groups, obtain a simple random sample from the groups, as a whole, and select all the individuals within the randomly selected group or stratum.

Point Estimation

Estimation stands as the first part of inferential statistics, while the second part is the hypothesis testing. There are two types of estimation: *point estimation* and *interval estimation*. We will address the point estimation in this section. The interval estimation will be addressed in the next section.

A point Estimate is that value of a statistic, which has been calculated from a sample, that estimates a parameter of the population.

Statistics as Estimators for parameters

It is clearly visible that we use statistics to estimate parameters due to the lack of time, energy, resources, and infinite populations. Statistics, from the sample, can be listed as: proportions, Arithmetic averages, ranges, quartiles, deciles, percentiles, variances, and standard deviations. It will become clear enough what each one means and what it will stand for.

The Sample Proportion

Suppose that there is a population in which each individual either does or does not have a certain characteristic. We like to consider a random sample out of this population. Selecting an individual from this population is an experiment that has ONLY two outcomes: either that individual has the characteristic we are interested in or does not. This kind of an experiment is called a Bernoulli experiment that has two outcomes, from now on labeled, a success or a failure. Let a random sample of size n be taken from this population. The sample proportion, denoted $\hat{p}$ by (read “p-hat”), is given by

$$\hat{p} = \frac{x}{n}$$

Here x is the number of individuals in the sample with the specified characteristic. The sample proportion ˆp is a statistic that estimates the population proportion, p, which is a parameter.

The Sample Mean and Sample Variance

Let there be a population with unknown mean "μ", (lower case Greek mu) and unknown standard deviation "σ" (lower case Greek sigma). To estimate μ and σ, we draw a simple random sample of size n: x_1, x_2, …,x_n. Then compute the sample mean

$$\overline{X} = 1/\mathrm{n} \sum_{i=1}^{n} X_i$$

and compute the sample variance

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

Then μ is estimated by the sample mean and σ is estimated by the sample variance.

Interval Estimation

An Interval Estimation is a range of values, calculated based on the information in the sample, that the parameter in a population will be within that range with some degree of confidence.

Suppose that a scientist wishes to weigh an ore sample with high precision. Because of a random error in measurements, he takes several readings and averages them by computing X. He may report his estimate of the true weight as:

1. As point estimate, X or

2. As an interval, X±E, where E stands for the error. It is the margin of error.

It is seen that a point estimate is a single value that is used to estimate an unknown population parameter. A point estimate is often insufficient because it is right or wrong. If the scientist reports that the estimated weight is 404 micrograms, (mg), then he does not really mean that the true weight is exactly 404 mg, and he should report how much error is involved. Also, it is insufficient to report that the true mean is within the range of values X±1S.E. because he does not report how much confidence he has in such a statement.

The best thing that can be done is to report estimates as “interval estimates”. An *interval estimate* (or a *confidence interval*) is a range of values used to estimate a population parameter with a certain degree of confidence. The confidence interval indicates the error of estimation in two ways: i) by the extent of its range and ii) by the probability of the true population parameter will be lying within that range.

In the following sections, we will introduce how to construct a confidence interval for estimating one or two population parameters, and specify the probability as a level of confidence.

Confidence Interval about One Parameter

In this section we will address the procedures on how to calculate the confidence interval on one parameter. Those parameters are the proportion and the mean.

Confidence Interval about One Proportion

It is of interest to estimate the proportion of employees, who favor a certain type of work, or the proportion of defective items in a certain lot, or the proportion of rats having a certain kind of symptoms. Let 'P' be the true proportion of elements that have attribute 'A' (a certain characteristic of interest) in a population. We draw a simple random sample of size n from this population and let 'X' equal number of elements in the sample that have attribute 'A'. Thus the point estimate of the true proportion 'p' in the population is given by ˆp = X/n, where X has a binomial distribution with parameters n and p. Recall that E(X) = np, and Var (X) = n.p. (1-p). Hence we find that E (ˆp) = p, and Var (ˆp) = p.(1-p)/n. By the Central Limit Theorem, the random variable given

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

has a standard normal distribution as 'n' increases. Thus

$$\mathrm{P}\,(-\ Z_{\mathrm{a}/2} < \mathrm{Z} < Z_{\mathrm{a}/2}\ ) = 1 - \alpha,$$

where Z is the value of the standard normal variable comprising a probability of alpha on its right. This with some algebra manipulations we reach the 100(1 – a) % C.I. (Confidence Interval) on p to be given by

$$\hat{p} - Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

Based on the above confidence interval, when the two limits are given, we can find the sample proportion by the following equation:

The lower limit + the upper limit = 2(sample proportion).

Confidence Interval about One Mean

Let there be a population with mean μ and variance σ^2. We wish to construct a confidence interval about, 'μ' with 100(1- a) % confidence level, where 0 < a < 1. There are three cases to be considered here. For the following cases, we will have a simple random sample of size n from the original population. Let that sample be X_i, i =1, 2, …, n.

Case I: The variance of the population is known, In this case the random variable

$$Z = \frac{\bar{X} - \mathrm{m}}{\mathrm{s}\ /\sqrt{n}}$$

has a N (0, 1) distribution and

$$\mathrm{P}\,(-\ Z_{\mathrm{a}/2} < \mathrm{Z} < Z_{\mathrm{a}/2}) = 1\ -\alpha.$$

With some algebra manipulations we reach the 100(1- a) % C. I. on 'μ' to be given by

$$\overline{x} - Z_{a/2}s \ /\sqrt{n} < \mathrm{m} < \overline{x} + Z_{a/2}s \ /\sqrt{n}$$

The above interval is called a z-interval about the mean μ. It is to be noticed here that we can find the sample mean or the margin of error when the limits of any confidence interval are given. This based on the following two equations:

The Lower limit + the Upper limit = 2(the sample mean), while

The upper limit – the lower limit = 2(the margin of error).

Case II: The variance of the population is unknown, with a large sample size (n ≥ 30)

Since 'σ' is not known, and we have a random sample, we estimate 'σ' by the standard deviation of the sample on hand. By replacing 'σ' by s in the above z-interval we reach at the following interval that will be the 100(1 – a) % C.I. on 'μ',

$$\overline{x} - Z_{a/2}S/\sqrt{n} < \mathrm{m} < \overline{x} + Z_{a/2}S/\sqrt{n}\,.$$

Sample Size Determination

Frequently, we wish to determine how large a sample should be in order to ensure that the error in estimating the population mean, or the population proportion, is less than a specified value of the error.

As it was shown in the derivation of the confidence interval for 'μ' and 'P', the margin of error was given by the following two formulas respectively

$$E = Z_{a/2}s \ /\sqrt{n}, \text{ and}$$

$$E = Z_{a/2}.\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

when the confidence level is taken to be 100(1 – a) % in both cases.

It is quite clear when the sample is too small; the required precision is not achieved. On the other hand, when the sample size is large, then some resources have been wasted. In order to meet the criterion of a specified margin of error calculations can be made to approximate the sample size needed in both cases of the mean and the proportion. Also, checking the formulas above for 'E', we find that each has 4 quantities. Thus whenever three of those are given the fourth can be found. In case of finding the sample size, the rounding will be up to the nearest whole number in order to meet the error criterion. Manipulating the above formulae we have

$$n=(Z_{\mathrm{a}/2}\mathrm{s}\ /E)^2.$$

This formula will give the sample size when E, the confidence level 100(1 – a) %, and the population standard deviation are given.

In case for finding the needed sample size to meet a certain criterion for estimating the population proportion, we have two cases to consider:

1) If there is an estimated value for that proportion p, and

2) If there is no information about P.

The formulae for n will be, respectively, given by

$$n=(Z_{\mathrm{a}/2}/E)^2.\hat{p}(1-\hat{p})\text{, and}$$

$$n=0.25.(Z_{\mathrm{a}/2}/E)^2.$$

Confidence Interval about Two Parameters

The confidence intervals to be calculated on two parameters will involve: a) two means, b) two proportions.

Confidence Interval about the difference between Two Proportions

Comparisons of proportions, in different groups, are a common practice. A whole-seller compares the proportions of defective items found in two separate sources of supply from which he buys these items. A safety engineer compares the proportions of head injuries sustained in an automobile accident by passengers with seat belts against those without seat belts.

Consider two independent samples of sizes n_1 and n_2 that are drawn from two binomial populations with parameters (i.e. probabilities of successes) p_1 and p_2. A 100(1 – a) % confidence interval will derived on the difference between p_1 and p_2 using the central limit theorem and the normal approximation to the binomial distribution.

Let x_1 and x_2 be the number of successes obtained in sample 1 and sample 2 respectively. We then have $\hat{}p_1 = x_1/n_1$, $\hat{}p_2 = x_2/n_2$ as the point estimates of p_1 and p_2 respectively. Moreover we have

$$\mathrm{E}\,(\hat{P}_1) = \hat{P}_1,\ \mathrm{E}\,(\hat{P}_2) = \mathrm{p}_2, \text{ with Var } (\hat{P}_1) = \mathrm{p}_1\,(1\text{- p1})/\mathrm{n1} \text{ and Var } (\hat{P}_2) = \mathrm{p}_2\,(1\text{- p}_2)/\mathrm{n}_2.$$

Because of the independence of the two samples, we can write

$$E(\hat{P}_1 - \hat{P}_2) = \mathrm{p}_1 - \mathrm{p}_2 \text{ with Var } (\hat{P}_1 - \hat{P}_2) = \mathrm{p}_1\,(1\text{- p}_1)/\mathrm{n}_1 + \mathrm{p}_2\,(1\text{- p}_2)/\mathrm{n}_2.$$

Now by the Central Limit Theorem, the random variable Z given by

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (\mathrm{p}_1 - \mathrm{p}_2)}{\sqrt{\mathrm{p}_1(1-\mathrm{p}_1)/\mathrm{n}_1 + \mathrm{p}_2(1-\mathrm{p}_2)/\mathrm{n}_2}}$$

has approximately a standard normal distribution with mean 0 and variance 1, i.e. $Z \approx N\,(0, 1)$. Hence

$$\mathrm{P}\,(-Z_{\alpha/2} < \mathrm{Z} < Z_{\alpha/2}) = 1 - \alpha,$$

where Z is as given above. With little algebra manipulations we reach at the 100(1- a) % confidence interval on the difference between the two proportions, $(p_1 - p_2)$ as given by the two limits

$$\text{Lower Limit: } (\hat{P}_1 - \hat{P}_2) - Z_{a/2}\sqrt{\hat{\mathrm{p}}_1(1-\hat{\mathrm{p}}_1)/\mathrm{n}_1 + \hat{\mathrm{p}}_2(1-\hat{\mathrm{p}}_2)/\mathrm{n}_2}\,,$$

$$\text{Upper Limit: } (\hat{P}_1 - \hat{P}_2) + Z_{a/2}\sqrt{\hat{\mathrm{p}}_1(1-\hat{\mathrm{p}}_1)/\mathrm{n}_1 + \hat{\mathrm{p}}_2(1-\hat{\mathrm{p}}_2)/\mathrm{n}_2}\,.$$

Note

The central limit theorem: Let $X_1,X_2, \ldots$ be any sequence of independent identically distributed random variables with finite positive variance. Let μ be the expected value and σ^2 the variance of each of the X_i. For $n \geq 1$, let Z_n be defined by

$$Z_n = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma};$$

then for any number a

$$\lim_{n \to \infty} F_{Z_n}(a) = \Phi(a),$$

where Φ is the distribution function of the $N(0,\ 1)$ distribution. In words: the distribution function of Z_n converges to the distribution function Φ of the standard normal distribution. Note that

$$Z_n = \frac{\bar{X}_n - \mathrm{E}\left[\bar{X}_n\right]}{\sqrt{\mathrm{Var}\left(\bar{X}_n\right)}},$$

Confidence Interval about difference between Two Means

It is quite often the following question is raised: Which average of those two means which are under investigation is better, or higher or smaller, or worse? In comparative experiments the investigator wishes to estimate the difference between two processes based on the difference between their means. For example, a chemist likes to compare the effects of two catalysts on the output of some chemical reactions. An agronomist wants to estimate the difference in yield of two varieties of corn. These questions and many others lead us to investigate how to estimate the difference between two means based on finding the confidence interval about that difference.

Assume that there are two populations with their means and variances given μ_i and σ^2_i for i= 1, 2 respectively. These populations could be normally distributed or not, as the discussion will reveal the cases below. We will select two simple random samples of sizes n_i, i = 1, 2, and denote them by X_j and Y_j, j = 1,2, …. n_i. Based on the data, from the samples, we can compute the mean and the variance for each sample, which are given by $\bar{X}$, S_1^2 and $\bar{Y}$, S_2^2 . There are two cases to be considered. Each case will be addressed based on the data and the information on hand.

Case I: The two population variances, σ^2_i for i = 1, 2, are known.

We know from earlier discussion that $\bar{X}$ and $\bar{Y}$ are normally distributed each has a normal distribution with mean and variance given by μ_1, σ_1^2/n_1 and μ_2 , σ_2^2/n_2 respectively, and thus the random variables

$$Z_1 = \frac{\bar{X} - \mathrm{m}_1}{\sqrt{\mathrm{s}_1^2 / n_1}}, \text{ and } Z_2 = \frac{\bar{Y} - \mathrm{m}_2}{\sqrt{\mathrm{s}_2^2 / n_2}},$$

each will have a standard normal distribution with mean 0 and variance 1. Because of the independence of the two samples we have $\bar{X} - \bar{Y}$ as the point estimate for $\mu_1 - \mu_2$ which it has a normal distribution with mean equal to $\mu_1 - \mu_2$ and variance given by $\sigma_1^2/n_1 + \sigma_2^2/n_2$ and thus the random variable

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mathrm{m}_1 - \mathrm{m}_2}{\sqrt{\mathrm{s}_1^2 / n_1 + \mathrm{s}_2^2 / n_2}}$$

Has a standard normal distribution with mean 0 and variance 1. Hence we have

$$P\ (-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1\ -\alpha$$

Now as it was the case when we derived the C.I. on the difference between two proportions, and with little algebra manipulation we reach at the 100(1 – a) % C.I. on the difference $\mu_1 - \mu_2$ to be given by the following two limits

$$\text{Lower Limit: } (\bar{X}\text{-}\bar{Y}) - Z_{\alpha/2}\sqrt{\mathrm{s}_1^2 / n_1 + \mathrm{s}_2^2 / n_2}$$

$$\text{Upper limit: } (\bar{X}\text{-}\bar{Y}) + Z_{\alpha/2}\sqrt{\mathrm{s}_1^2 / n_1 + \mathrm{s}_2^2 / n_2}\ .$$

In other words, the 100(1 – a) % C.I. on the difference $\mu_1 - \mu_2$ is given by

$$(\bar{X}\text{- }\bar{Y})\ \text{-}Z_{\alpha/2}\sqrt{\mathrm{s}_1^2 / n_1 + \mathrm{s}_2^2 / n_2} < \mathrm{m}_1 - \mathrm{m}_2 < (\bar{X}\text{- }\bar{Y})\ +Z_{\alpha/2}\sqrt{\mathrm{s}_1^2 / n_1 + \mathrm{s}_2^2 / n_2}\ .$$

Case II The two population variances S_i^2 for i = 1, 2 , are unknown, Large Sample Size

In this case the question that will be asked is: What are the sample sizes? For sample sizes of greater than 30 each, by using the Central Limit Theorem, and replacing the population variances by their estimates, from the sample we see that the random variable given by

$$Z = \frac{(\bar{X} - \bar{Y}) - (m_1 - m_2)}{\sqrt{S_1^2 / n_1 + S_2^2 / n_2}},$$

has a standard normal distribution. Hence the 100(1 – a) % C.I. on the difference $\mu_1 - \mu_2$ will be give by

$$(\bar{X}\text{-}\bar{Y}) - Z_{a/2} \sqrt{S_1^2 / n_1 + S_2^2 / n_2} < m_1 - m_2 < (\bar{X}\text{- }\bar{Y}) + Z_{a/2} \sqrt{S_1^2 / n_1 + S_2^2 / n_2} .$$