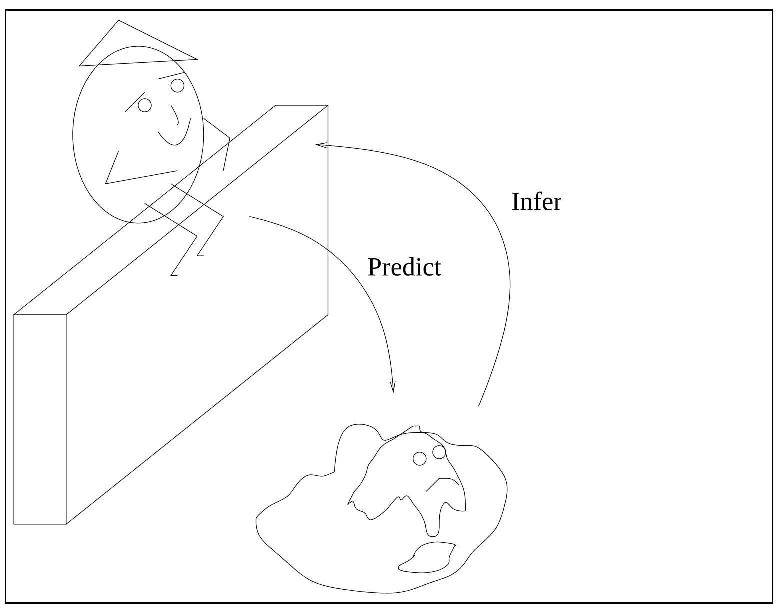
Probability for STAT112

Charles Fleming

February 20, 2019

1 Probability



Humpty Dumpty sat on a wall
Humpty Dumpty had a great fall
All the King's horses and all the King's men
Could not put Humpty Dumpty together again.

Probability and statistics go together like ham and eggs. They are two sides of the same coin. Probability goes in one direction; statistics goes in the other, in the same sense as differential and integral calculus, or in the sense of Humpty Dumpty's ill starred fate. A probabilist is concerned with predicting what will Humpty Dumpty look like after his fall. *All the King's horses and all the King's men* are the statisticians who try to infer based on the gory details what Humpty Dumpty looked like before his fateful fall. The probabilist tells us the laws and theorems which are used to predict the outcome; the statisticians use those laws and theorems to infer the origins of the data.

The usefulness of descriptive statistics is limited to that purpose of describing a set of data by a few numbers. To go beyond the scope of descriptive statistics, it is necessary to develop a new set of tools drawn from the science of probability. To that end, we will use the theorems and procedures that are employed in probability for predicting the outcome of an event.

Given a set of initial conditions and a model which describes the evolution of a phenomenon, the probabilist tries to predict the outcome of an event. For example, given the initial conditions of a bow and arrow and the equations of motion of a projectile, the probabilist will seek to determine the probability that the arrow will hit the bull's-eye. The statistician, on the other hand, is given the impaled target with arrows and wants to know where the arrows came from. He will use the same equations of motion and laws of probability to work backwards from the data to a description of the population in a process called inference.

In order to understand the techniques and concepts which the statistician utilizes, it is necessary to study probability. In order to understand probability, it is necessary to study abstract mathematics, because ironically abstract mathematics makes the concepts of probability easy to understand. The key idea underlying probability is the notion of size. In essence, the size of an event is called a probability; therefore, the study of probability begins with the theory of sets and with learning different ways of measuring their sizes. This is 20^{th} century mathematics which Emile Borel and Henri Lebesgue were instrumental in developing. It is abstract mathematics at its abstract; it is quite advanced and goes by the name of measure theory. Although the mathematics of statistics becomes very sophisticated very quickly, we will take advantage of the heuristic notions of measure theory to understand the foundations of probability, and, in the same vein, we will not dwell on the usually difficult mathematical derivations of the numerous formulas which we are about to study.

2 Operations on Sets

We chose $\mathcal{P}$ to denote the set of elements of a population, and we chose $\mathcal{S}$ to denote the set of elements of a sample which are drawn from a list $\mathcal{L}$ consisting of names for the purpose of identifying elements of the population. From the beginning of the study of statistics, the essential elements of our interest are collected into sets. Likewise, our study of probability begins with the essential properties of sets and will eventually end with the formulation of the basic though indispensable mathematical tools of the statistical trade. To arrive at that end, we need to commence the study of probability by looking at a set.

Let Ω be a set.

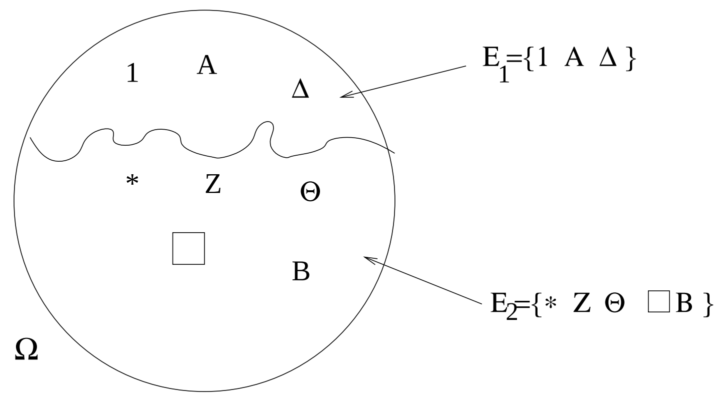


Figure 1

Definition 1. *If $A \subseteq \Omega$ and $B \subseteq \Omega$, then the set of all elements which are common to both A and B is called the **intersection** of A and B. It is denoted by $A \cap B$.*

Definition 2. *If $A \subseteq \Omega$ and $B \subseteq \Omega$, then the set of all elements of either A or B with no duplications is called the **union** of A and B. It is denoted by $A \cup B$.*

Definition 3. *If $A \subseteq \Omega$ and $B \subseteq \Omega$ and $A \cap B$ is empty, then A and B are called **disjoint** sets. The empty set is denoted by $\emptyset$.*

For example, let $A = \{1, 3, \alpha, w, -1\}$ and let $B = \{\alpha, \beta, -1, 0, e\}$. The union of A and B is $A \cup B = \{\alpha, \beta, e, w, -1, 0, 1, 3\}$. By convention, we omit duplications. The intersection of A and B is $A \cap B = \{\alpha, -1\}$.

Suppose Ω consists of eight elements as shown in Figure 1 and that Ω is divided into two partitions, E_1 and E_2. The size of $\Omega = 8$; the size of $E_1 = 3$, and size of $E_2 = 5$. Equivalently, the sizes of each set can be reported relative to the size of Ω, so that

$$\frac{size\ of\ E_1}{size\ of\ \Omega} = \frac{3}{8} \quad \frac{size\ of\ E_2}{size\ of\ \Omega} = \frac{5}{8} \quad \frac{size\ of\ \Omega}{size\ of\ \Omega} = \frac{8}{8}$$

Rather than write the phrase, $\frac{size\ of}{size\ of\ \Omega}$, over and over again, we will substitute in its place a certain set function according to the following definition:

Definition 4. *Define $P()$ such that $P(E) = \frac{size\ of\ E}{size\ of\ \Omega}$.*

In our example then, $P(E_1) = \frac{3}{8}$, $P(E_2) = \frac{5}{8}$, and $P(\Omega) = \frac{8}{8}$.

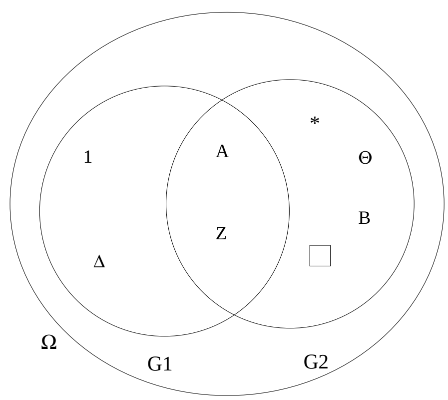
Because E_1 and E_2 have no common elements, they are disjoint; therefore, $E_1 \cap E_2 = \emptyset$, and the union of E_1 and E_2 consists of all the elements of Ω; therefore, $E_1 \cup E_2 = \Omega$. A disjoint partitioning of Ω has a nice property with regard to measuring its size in terms of the sizes of its constituents. Observe that $P(E_1)+P(E_2) = \frac{3}{8}+\frac{5}{8} = 1 = P(\Omega) = P(E_1 \cup E_2)$ i.e. $P(E_1 \cup E_2)$=$P(E_1) + P(E_2)$. With that observation in mind, we arrive at a very important theorem:

Theorem 1. *If E_1 and E_2 are subsets of Ω and if they are disjoint, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.*

Sometimes, when working with sets, our attention is focused on only one of them while the rest are put together into the complement.

Definition 5. *The **complement** of E, denoted by E^c, is the set of all elements of Ω which are not elements of E.*

Corollary 1. *$E \cup E^c = \Omega$.*



Rather than partition Ω into two disjoint sets, E_1 and E_2, suppose two other subsets had been defined like: $G_1 = \{1\ A\ \Delta\ Z\}$ and $G_2 = \{*\ Z\ \theta\ \square\ B\ A\}$. Although their union is Ω, i.e. $G_1 \cup G_2 = \{1\ Z\ A\ \Delta\ *\ \theta\ \square\ B\} = \Omega$, G_1 and G_2 are not disjoint because they have common elements: $G_1 \cap G_2 = \{A\ Z\}$.

Consider measuring their relative sizes: $P(G_1) = \frac{4}{8}$, $P(G_2) = \frac{6}{8}$, and $P(G_1 \cup G_2) = P(\{1\ A\ \Delta\ Z\ *\ \Theta\ \square\ B\}) = P(\Omega) = 1$. But notice that $P(G_1) + P(G_2) = \frac{4}{8} + \frac{6}{8} = \frac{10}{8} \neq 1 = P(\Omega) = P(G_1 \cup G_2)$. The size of the union of G_1 and G_2 is not equal to the sum of their sizes, if we blindly believed Theorem 1. Theorem 1 does not apply because its condition is not satisfied by the construction of G_1 and G_2. Theorem 1 is valid only for disjoint sets; however, $G_1 \cap G_2 \neq \varnothing$. In order to rectify Theorem 1 for general application, it is sufficient to observe that $G_1 \cap G_2$ is counted twice when Theorem 1 is used, once when G_1 is measured and again when G_2 is measured. By taking away one count of $G_1 \cap G_2$, we produce a general theorem.

Theorem 2. *If G_1 and G_2 are subsets of Ω, then $P(G_1 \cup G_2) = P(G_1) + P(G_2) - P(G_1 \cap G_2)$.*

Proof. To prove this theorem, we will use the trick that $G_1 \cup G_2 = G_1 \cup (G_2 - G_1) = G_1 \cup G_1^c \cap G_2$.

By Theorem 1, $P(G_1 \cup G_2) = P(G_1 \cup G_1^c \cap G_) = P(G_1) + P(G_1^c \cap G_2)$ because G_1 and $G_1^c \cap G_2$ are disjoint. We add zero to the right hand side, so that, $P(G_1 \cup G_2) = P(G_1) + P(G_1^c \cap G_2) + P(G_1 \cap G_2) - P(G_1 \cap G_2)$ and use Theorem 1 again to combine $P(G_1^c \cap G_2) + P(G_1 \cap G_2)$ into $P(G_1^c \cap G_2 \cup G_1 \cap G_2)$ which is equal to $P((G_1^c \cup G_1) \cap G_2)$. Therefore, $P(G_1 \cup G_2) = P(G_1) + P(\Omega \cap G_2) - P(G_1 \cap G_2) = P(G_1) + P(G_2) - P(G_1 \cap G_2)$. ■

Definition 6. *A complete listing of all subsets of a set Ω is called the **power set** of Ω.*

Example 1. *If $\Omega = \{1\ A\ \Delta\ *\}$, then*

$$\textit{the power set of } \Omega = \left\{ \begin{array}{llllll} \{\} \\ \{1\} & \{A\} & \{\Delta\} & \{*\} \\ \{1A\} & \{1\Delta\} & \{1*\} & \{A\Delta\} & \{A*\} & \{\Delta*\} \\ \{A\Delta*\} & \{1\Delta*\} & \{1A*\} & \{1A\Delta\} \\ \{1A\Delta*\} \end{array} \right\}$$

If a theorem can be proven for a power set, then it is true for all members of the power set. The concept of a power set is useful when developing the theory of probability and statistics, in order to guarantee complete generality and to prevent any exceptions to the theorems to exist. A good mathematician prides himself in developing a theory which is watertight and is certain not to have any unaccounted exceptions. Everything in mathematics must be consistent and complete. The same motivation applies to probability. In that light, a casual reflection of the definition of P() will reveal a major flaw in that the function, P(), could have been defined to give any arbitrary value like: $P(G_1) = 0 \quad P(G_2) = 1$; or $P(G_1) = \frac{1}{2} \quad P(G_2) = \frac{1}{2}$; or $P(G_1) = \frac{9}{10} \quad P(G_2) = \frac{1}{10}$. There is nothing mentioned thus far which dictates the definitive value of P() when it is applied to a set. The function, $P()$, has been arbitrarily defined, a circumstance which is not satisfactory for a mathematician and begs for the establishment of a solid foundation upon which to build the calculus of probability.

Blaise Pascal
1623-1662

Pierre de Fermat
1601-1665

Historians of mathematics seem to agree that the birth of mathematical probability occurred in 1654 during the correspondence of Pascal and Fermat. Not until 1933 was probability placed on a solid foundation by the great 20^{th} century mathematician, Andrei Kolmogorov.

Andrei Nikolaevich Kolmogorov
(Андрей Николаевич Колмогоров)
1903-1987

3 Formal Definition of Probability

Axiom 1. $0 \leqslant P(E) \leqslant 1$.

Axiom 2. $P(\Omega) = 1$.

Axiom 3. *If $E_1,\ E_2, \cdots, E_n$ are pairwise disjoint subsets of Ω, then $P(E_1 \cup E_2 \cup \cdots \cup E_n) = P(E_1) + P(E_2) + \cdots + P(E_n)$.*

Definition 7. *A function, P, that satisfies Axioms 1–3 is called a* ***probability****.*

An axiom is a fundamental statement which cannot be proven. The axioms of probability support all the formulas which we will use, but the axioms are not sufficient. There is nothing in the definition of probability which tells us what a probability should be in a given situation. The missing piece of the puzzle must correspond to a fundamental characteristic of the phenomenon which is being studied.

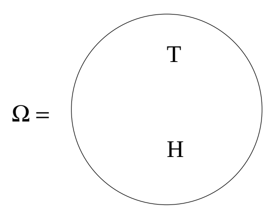
Before we proceed with the development of probability, the following definitions will make our discussion easier.

Definition 8.

> *The* ***sample space****, Ω, is a set which consists of all possible outcomes.*
>
> *An element of the sample space, Ω, is called an* ***outcome****.*
>
> *A subset, E, of the sample space, Ω, is called an* ***event****.*
>
> *If the event, E, consists of only one element, then E is called a* ***simple event****.*

Let us conduct a simple experiment to illustrate the concept of probability. The experiment is abstract, meaning that it will occur in our imagination. In this experiment, one coin will be flipped. The outcome of getting a tail will be denoted by T, and the outcome of getting a head will be denoted by H.

The sample space consists of two elements as depicted here.



Let $E_1 = \{T\}$ and $E_2 = \{H\}$. E_1 and E_2 are simple events; they are disjoint; their union comprises the sample space. The fundamental characteristic of the experiment which will dictate the value of a probability is the stipulation that the outcomes are equally likely to occur. Assume that it is equally likely that the outcome T occurs as the outcome H, i.e. $P(E_1) = P(E_2) = p$. Let $P()$ be a probability which implies that, by Axioms 1–3, $P(\Omega) = 1$. Because $E_1 \cap E_2 = \varnothing$, E_1 and E_2 are disjoint and, by Theorem 1, $1 = P(\Omega) = P(E_1 \cup E_2) = P(E_1) + P(E_2) = p + p = 2p$; therefore, $p = \frac{1}{2}$.

Recall that according to the original definition of P(), $P(E_1) = \frac{\textit{size of } E_1}{\textit{size of } \Omega} = \frac{1}{2}$. It appears that the original definition of P() corresponds to the probability of equally likely outcomes. The notion of equally likely outcomes and the meaning of the phrase a *fair coin* or *fair dice* are the same. If the outcomes are not equally likely, then we need more information about the underlying phenomenon, in order to determine the right value of a probability.

Let us conduct another experiment in which two fair coins are tossed.

Example 2. *Two fair coins are tossed. The sample space is:*

$\Omega =$

T T
T H
H T
H H

All possible outcomes are $\{TT \quad TH \quad HT \quad HH\}$. *In general, n fair coins produce* 2^n *possible outcomes. Let* $E = \{T\ T\}$. *Find P(E)=P(event that both coins will land tails up)=p. We have at our disposal two methods of finding the probability that the event, E, will occur.*

Method I $p = P(E) = \frac{size\ of\ E}{size\ of\ \Omega} = \frac{1}{4}$

Method II

$$\begin{aligned} 1 &= P(\Omega) = P(E) + P(\{T\ H\}) + P(\{H\ T\}) + P(\{H\ H\}) \\ 1 &= p + p + p + p = 4p \quad \textit{by the assumption of equally likely outcomes} \\ 1 &= 4p \rightarrow p = \frac{1}{4} \end{aligned}$$

It should not be forgotten that a probability is the relative measure of the size of an event to the sample space.

Example 3. *Let E be the event that at least one tail appears. Therefore,* $E = \{TT \quad TH \quad HT\}$.
P(E)=P(event that at least one tail appears)=$\frac{size\ of\ E}{size\ of\ \Omega} = \frac{3}{4}$

Example 4. *Let G be the event that at most one tail appears. Therefore,* $G = \{HH \quad HT \quad TH\}$.
P(G)=P(event that at most one tail appears)=$\frac{size\ of\ G}{size\ of\ \Omega} = \frac{3}{4}$

As the complexity of the problems grow, it will be helpful to have tools to use in solving them. For example, the probability of the complement is useful to know.

Theorem 3. $P(E^c) = 1 - P(E)$.

Proof. We note that since E and E^c are disjoint and that $E \cup E^c = \Omega$, we may use Theorem 1; therefore,

$$\begin{array}{ll} P(E \cup E^c) & = P(E) + P(E^c) \\ \quad\| & \\ P(\Omega) & = 1 \to P(E^c) = 1 - P(E) \end{array}$$

■

We may enumerate all possible outcomes for measuring the size of E^c, or we may use Theorem 3 to obtain the same answer.

Example 5. *Suppose* $E = \{TT \quad TH \quad HT\}$, *then* $E^c = \{HH\}$ *hence* $P(E^c) = \frac{size\ of\ E^c}{size\ of\ \Omega} = \frac{1}{4}$. *Or using Theorem 3 and the answer of the previous example,* $P(E^c) = 1 - P(E) = 1 - \frac{3}{4} = \frac{1}{4}$.

In general. when the outcomes are equally likely or, in other words, in fair games: $P(E) = \frac{size\ of\ E}{size\ of\ \Omega} = \frac{number\ of\ possibilities}{total\ number\ of\ outcomes}$.

Consider a more complicated example in which a fair coin is tossed and a fair die is rolled. The sample space consists of all possible outcomes in which T will denote a tail, H will denote a head, and the numbers 1, 2, 3, 4, 5, and 6 will denote the faces of the die.

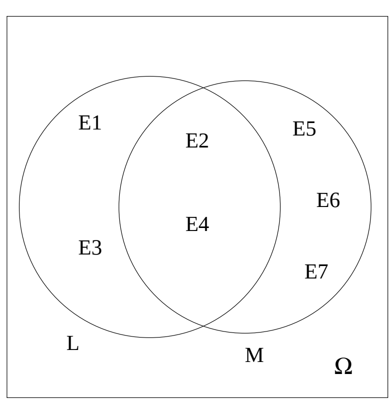
Example 6. *Flip a fair coin and toss a fair four-sided die.*

$\Omega = \{T1 \ \ T2 \ \ T3 \ \ T4 \ \ H1 \ \ H2 \ \ H3 \ \ H4\}$.

Let E be the event of getting at most a 3. Then $E = \{T1 \quad T2 \quad T3 \quad H1 \quad H2 \quad H3\}$; *therefore,* $P(E) = \frac{6}{8}$.

Let G be the event of getting a tail and at least a 2, then $G = \{T2 \ \ T3 \ \ T4\}$; *therefore,* $P(G) = \frac{3}{8}$.

Suppose that the events are not equally likely to occur. Sufficient information must be provided, in order to make it possible to find the probabilities of the events. In the following example, the probability of every simple event is given in the statement of the problem.



Example 7. $P(E_1) = P(E_5) = \frac{1}{20}$ $P(E_2) = P(E_4) = \frac{1}{10}$ $P(E_3) = P(E_7) = \frac{1}{5}$ *and* $P(E_6) = \frac{3}{10}$.

Find $P(L)$*:*

ANS: $P(L) = P(E_1 \cup E_2 \cup E_3 \cup E_4) = P(E_1) + P(E_2) + P(E_3) + P(E_4) = \frac{1}{20} + \frac{1}{10} + \frac{1}{5} + \frac{1}{10} = \frac{9}{20}$. *We use Theorem 1 because all simple events are mutually disjoint.*

Find $P(L \cap M^c)$*:*

ANS: $P(L \cap M^c) = P(E_1 \cup E_3) = P(E_1) + P(E_3) = \frac{1}{5} + \frac{1}{20} = \frac{1}{4}$

Consider the experiment of rolling two fair dice. The sample space consists of all 36 possible outcomes:

$$\Omega = \left\{ \begin{array}{cccccc} 11 & 12 & 13 & 14 & 15 & 16 \\ 21 & 22 & 23 & 24 & 25 & 26 \\ 31 & 32 & 33 & 34 & 35 & 36 \\ 41 & 42 & 43 & 44 & 45 & 46 \\ 51 & 52 & 53 & 54 & 55 & 56 \\ 61 & 62 & 63 & 64 & 65 & 66 \end{array} \right\}$$

What might look like numbers are actually names of the outcomes. The first digit corresponds to the face of the first die, and the second digit corresponds to the face of the second die. The order of the digits is important. Imagine that the first die is colored red and the second one colored white. The two dice are distinct, so that the outcome, 12, is different from the outcome, 21. The order of faces is essential in this example.

Problem 1. *What is the probability of the event, S, of casting sevens?*

By drawing outcomes from out of the sample space, we construct the event

$$S = \{16, \quad 25, \quad 34, \quad 43, \quad 52, \quad 61\}$$

Because the outcomes are equally likely by the assumption that the dice are fair,

$$P(S) = \frac{number\ of\ possibilities\ comprising\ the\ event}{total\ number\ of\ outcomes} = \frac{6}{36} = \frac{1}{6}$$

Likewise for the event, E, of casting an even number, the appropriate outcomes are first listed. Let E be the event of casting an even number; therefore,
$E = \{11, \ 13, \ 15, \cdots, 62, \ 64, \ 66\}$, so that $P(E) = \frac{18}{36} = \frac{1}{2}$.

What is the probability of casting a seven and getting an even number? Obviously, such an event is impossible because seven is an odd number. To answer the question formally, we find that $P(S\ and\ E) = P(S \cap E) = P(\varnothing) = \frac{0}{36}$.

The probability of the event of casting a seven or getting an even number can be found either by finding the relative size of the event once its outcomes have been listed or the probability can be found by means of Theorem 1 since the events S and E are disjoint, so that $P(S\ or\ E) = P(S \cup E) = P(S) + P(E) = \frac{6}{36} + \frac{18}{36} = \frac{2}{3}$.

Let F be the event of rolling a 4 on any face, then

$$F = \{14,\ \ 24,\ \ 34,\ \ 44,\ \ 54,\ \ 64,\ \ 41,\ \ 42,\ \ 43,\ \ 45,\ \ 46\}$$

Because of the rule which stipulates that duplicates are omitted allows only one 44 is listed instead of two 44's; therefore, $P(F) = \frac{11}{36}$. The event of casting a seven and getting a four on any face is the same as $S \cap F = \{43,\ \ 34\}$, so that $P(S\ and\ F) = P(S \cap F) = \frac{2}{36} = \frac{1}{18}$.

Example 8. *What is the probability that the sum of the faces will be at most a 9 when two fair dice are cast?*

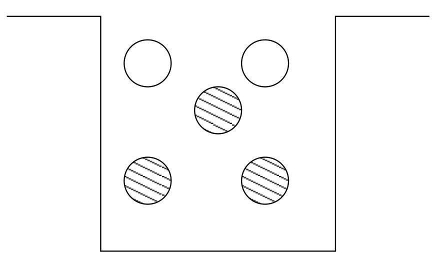
1. *Let E be the event that the sum of the faces is at most a 9. A listing of E is:*

$$E = \left\{ \begin{array}{llllll} 11 & 12 & 13 & 14 & 15 & 16 \\ 21 & 22 & 23 & 24 & 25 & 26 \\ 31 & 32 & 33 & 34 & 35 & 36 \\ 41 & 42 & 43 & 44 & 45 & \\ 51 & 52 & 53 & 54 & & \\ 61 & 62 & 63 & & & \end{array} \right\}$$

2. *Since the dice are fair,* $P(E) = \frac{size\ of\ E}{size\ of\ \Omega} = \frac{30}{36} = \frac{5}{6}$.

Another very popular experiment from which a vast array of problems in science ranging from genetics to nuclear reactions can be modeled is the one of drawing balls from an urn of different colored balls.

Example 9. *Urn*



Suppose the urn contains 10 red balls and 7 white balls. What is the probability that a ball drawn at random is a white ball?

1. *Find the sample space:*

$$\Omega = \{R\ \ R\ \ R\ \ R\ \ R\ \ R\ \ R\ \ R\ \ R\ \ R\ \ W\ \ W\ \ W\ \ W\ \ W\ \ W\ \ W\}$$

2. *Define E to be the event that a ball is white; therefore,*

$$E = \{W \ \ W \ \ W \ \ W \ \ W \ \ W \ \ W\}$$

and $P(E) = \frac{size\ of\ E}{size\ of\ \Omega} = \frac{7}{17}$.

Problems in which random sampling occurs raises a very important question. What is meant by the phrase: draw at random? Referring to the previous example, the process of picking a white ball is a random process only if $\frac{\#\ of\ white\ balls}{total\ \#\ of\ balls} \rightarrow \frac{7}{17}$. If the object is a fair coin, the process is random only if $\frac{\#\ of\ heads}{total\ \#\ of\ tosses} \rightarrow \frac{1}{2}$. If a die, then $\frac{\#\ of\ occurrences\ of\ a\ face}{total\ \#\ of\ casts} \rightarrow \frac{1}{6}$. The arrows signify that the number of drawings tends to infinity, so that the ratio converges in the limit to the theoretical value for a fair process. If the concept of drawing something at random seems abstruse, it is because it is abstruse. The notion of randomness has provoked mathematicians and logicians to make attempts to formulate a rigorous definition of randomness with the thought in mind that any pattern whatsoever in the series of outcomes opposes the concept of randomness. Based on intuition, a process is deemed random if it is impossible to predict the next outcome. Drawing a certain colored ball from an urn should be absolutely unpredictable, if the drawing is performed at random. Implementing this intuitive notion of randomness in a real experiment constitutes the conundrum of the problem. As soon as a method is proposed to produce a truly random sequence of outcomes like random digits, an exception is discovered which shows that the process is either not feasible or not truly random. It is in the imaginary world of probability that a truly random sampling can be implemented. In practice, there is no known mechanism to draw a sample at random. Yet with each attempt of giving a definitive meaning of randomness, mathematicians are coming closer to a final answer which appears not to be too far in the future. The notion of randomness is one of those concepts whose subtleties belie its simplicity.

4 Conditional Probability

There might be given additional information about an experiment which could improve the knowledge of the likelihood of an outcome.

Suppose we know already that sevens have been cast after rolling a pair of fair dice. What is the probability of getting a 4 on a face given that sevens have been cast? If it is known for certain that a seven has been cast, then the sample space of all possible outcomes must be:

$$\Omega = \left\{ \begin{array}{cccccc} & & & & & 16 \\ & & & & 25 & \\ & & & 34 & & \\ & & 43 & & & \\ & 52 & & & & \\ 61 & & & & & \end{array} \right\}$$

The event of getting a four on any face given that a seven has been cast must then be: $E = \{34, \;\; 43\}$, so that given the additional information about the experimental outcomes, the size of the sample space decreases with the consequence that

$$\begin{aligned} P(getting\ a\ four\ given\ that\ sevens\ were\ cast) &= \frac{number\ of\ possibilities}{total\ number\ of\ outcomes} \\ &= \frac{2}{6} = \frac{1}{3} = \frac{12}{36} \end{aligned}$$

The conditional probability reduces the size of the sample space from 36 to 6 with the consequence that the probability increases from $\frac{11}{36}$ given on page 11 to $\frac{12}{36}$ given above.

Definition 9. *Denote the* ***conditional probability*** *by $P(A|B)$ which means: the probability of the event, A, given that the event, B, has occurred. Also $P(A|B) = \frac{P(A \cap B)}{P(B)}$.*

Example 10. *From page 11, $P(F \cap S) = \frac{2}{36}$ and $P(S) = \frac{1}{6}$; therefore,*

$$\begin{aligned} P(F|S) &= P(getting\ a\ four\ given\ that\ sevens\ were\ cast) \\ &= \frac{P(F \cap S)}{P(S)} = \frac{\frac{2}{36}}{\frac{1}{6}} = \frac{12}{36} = \frac{1}{3} \end{aligned}$$

which is in agreement with the above calculation gotten by direct enumeration of the event.

In general, $P(A) \neq P(A|B)$ as in $P(F) = \frac{11}{36} \neq \frac{1}{3} = P(F|S)$. Not all information is useful, however. Suppose the problem was posed this way: what is the probability of getting a four on any face when two fair dice are cast given that the New York Yankees won the World Series in 1927? Whether or not the Yankees ever won the World Series is irrelevant to the casting of fair dice. In other words, the event, B, defined by the Yankees winning the World Series does not influence the outcome of the event, A. In mathematical terms, the statement which describes the event of casting a four on any face given that the Yankees won the World Series in 1927 is written as: $P(A|B) = P(A)$. In other words, the probability of casting a four on any face of a pair of fair dice is independent of the history of baseball. The probability of A does not change because the event, B, has no influence on the outcome of event, A. The independence of an outcome on another event is usually very desirable in statistics.

Definition 10. *If $P(A|B) = P(A)$, then the events A and B are said to be* ***independent****.*

The concept of independence plays a central role in statistics. If there were no such condition as independence, statistics would be essentially impractical for ordinary use. That is why a statistician takes great pains to demonstrate that the events are independent from each other; otherwise, the statistical analysis will be too complex and too formidable to reach even a simple conclusion.

An alternative expression of independence is given by the following theorem:

Theorem 4. *Suppose that events A and B are independent, then $P(A \cap B) = P(A)P(B)$.*

Proof. By independence,

$$\begin{array}{cc} P(A|B) & = P(A) \\ \| & \\ \dfrac{P(A \cap B)}{P(B)} & \end{array}$$

Therefore, by cross multiplying: $P(A \cap B) = P(A)P(B)$. ■

Conversely,

Corollary 2. *The events, A and B, are said to be independent if $P(A \cap B) = P(A)P(B)$.*

The concept of independent events bears no relation to the concept of disjoint sets. Independent events are defined only in terms of probability whereas sets having no common elements are disjoint sets without any mention of probability whatsoever.

Problem 2. *A committee of quality control engineers at Westinghouse Electric Corporation evaluated the judgment of inspectors in assessing the quality of 153 soldered joints. A tabulation of the results appears below.*

		Inspectors		
		Acceptable	Defective	Total
Committee	Acceptable	101	10	111
	Defective	23	19	42
	Total	124	29	153

a. *Let E be the event that the inspectors determine a joint to be acceptable. Out of 153 cases which comprise the sample space, the inspectors accepted 124 of them. Assuming that the outcomes are equally likely to occur, $P(E) = \frac{124}{153}$. On the other hand, the probability that the inspectors rejected a soldered joint is the probability of the complement that they accepted it; that is, $P(E^c) = 1 - \frac{124}{153} = \frac{29}{153}$.*

   The counterpart to the group of inspectors is the committee of experts. Let G be the event that the committee determines a joint to be acceptable. Using the same reasoning as above for finding the probabilities about the inspectors, the probability that the committee accepted a soldered joint is: $P(G) = \frac{111}{153}$.

b. *The probability that the committee and the inspectors agree that a joint is good is: $P(E \cap G) = \frac{101}{153}$. The probability of the complement is the probability that both cannot accept a joint, so that $P((E \cap G)^c) = 1 - \frac{101}{153} = \frac{52}{153}$. The interpretation of the complement implies that the committee and the inspectors not only might have agreed to reject a joint but on some occasions they might have disagreed; that is, one might have accepted a joint while the other rejected it.*

c. *Let H be the event that the committee and the inspectors simply agree one way or the other, then $P(H) = \frac{101+19}{153} = \frac{120}{153}$. The probability that they disagree is: $P(H^c) = 1 - \frac{120}{153} = \frac{33}{153}$. The event that both groups agree includes the event that both reject a joint whereas in the previous section both groups are seen to agree only when a joint is a good one.*

d. *What is the probability that the committee will determine a joint to be acceptable given that the inspectors have already accepted the joint? $P(G|E) = \frac{P(E \cap G)}{P(E)} = \frac{\frac{101}{153}}{\frac{124}{153}} = \frac{101}{124}$. The purpose of asking this question is to see if there is a dependency of the committee on the judgment of the inspectors. If everyone is judging the joints correctly, then the committee should agree with the decisions of the inspectors; $P(G|E)$ should be 1.*

   From part a, $P(G) = \frac{111}{153}$, but $P(G|E) = \frac{101}{124}$; therefore, G and E cannot be independent events. The results of Westinghouse's experiment should cause one to wonder why the events, E and G, are not independent. The committee and the inspectors apparently are not acting independently. Perhaps some members of the committee and some inspectors are acting in collusion with one another, or members of both groups accurately followed the same soldering standards, or the experiment was flawed somehow.

e. *What is the probability that the committee will judge a joint acceptable when the inspectors have rejected it? $P(G|E^c) = \frac{P(G \cap E^c)}{P(E^c)} = \frac{\frac{10}{153}}{\frac{29}{153}} = \frac{10}{29}$. But this should be 0, if every one is doing his job correctly. Instead the probability is almost 35 percent. The management at Westinghouse could have, for all practical purposes, flipped a coin* *in deciding whether or not the inspectors' decisions in rejecting joints were valid or invalid. In conclusion, perhaps the inspectors need to receive more training in assessing soldered joint, or a better experiment needs to be performed to give more reliable data.*

A more important question is raised by this problem: Do the probabilities reflect actual events? The answer to this question may be yes or no. In any case, the question is not well formed in the first place. It is a trick question. In the axioms of probability or in any example in which the value of a probability has been deduced, the sample space is the starting point. A sample space and a population are fundamentally different. A population consists by our definition of a set of real things which have substance and which are perceptible. A sample space is a set of outcomes of an imaginary experiment. The outcomes exist only in our imagination. They have no substance; they cannot be perceived by any sense or instrument. A probability of an event is the relative size of that subset of outcomes to the size of the sample space. Thus we talk about casting dice or flipping coins only in the context of an imaginary experiment. To say that the probability of getting a head is 1/2 bears no relation to the result of an actual flipping of a real coin. A probability is defined in terms of sets whereas in a real experiment the outcome is either a head or a tail, regardless of the notion of probability. An imaginary experiment and a real experiment are not the same.

The axioms of probability purport no connection with the real world, whereas a statistician deals with observations of the real world. The world of statistics and the world of probability are separate and different. One is real; the other is imaginary. Later, we will endeavor to build a bridge between these two worlds, and, by that bridge, we will introduce into the world of statistics the full power of our imagination which will wield the vast array of mathematical tools which will be at our disposal.

It should come as no surprise that problems found in the study of probability can be very complex and very difficult to solve. Predicting the state of the national economy or the weather a few weeks in advance are two such notoriously difficult problems. Successful formulations of models and the associated mathematical theories for manipulating them bring Nobel prizes to economists. A common tactic in solving a complex problem is the military tactic of divide and conquer. Deducing the probability of a complicated event might be possible, if the event is broken into manageable pieces. The process by which an event is decomposed does not follow a methodical recipe, rather it relies on ingenuity. An indispensable tool for conquering a complicated event is conditional probability. It can be used successively many times until the event is adequately decomposed to render it solvable. Ingenuity enters the process of conceiving a particular ancillary event, E, upon which the original event, A, is conditioned. Both this ancillary event and its complement are used for the purpose of reconstituting the original event from the pieces. If the ancillary event, E, is very cleverly conceived, then it will be easy to deduce the probabilities of

$P(A|E)$ and $P(A|E^c)$. Conditional probability forms the essence of the method of breaking an event into pieces and the mechanism of doing so is expressed by Theorem 5.

Theorem 5. *Let A and E be events, then* $P(A) = P(A|E)P(E) + P(A|E^c)P(E^c)$.

This theorem decomposes an event into simpler ones. It should be noted that the decomposition is done by means of an ancillary event, E, which is deliberately invented for the purpose of breaking the event, A, into smaller and presumably more manageable pieces.

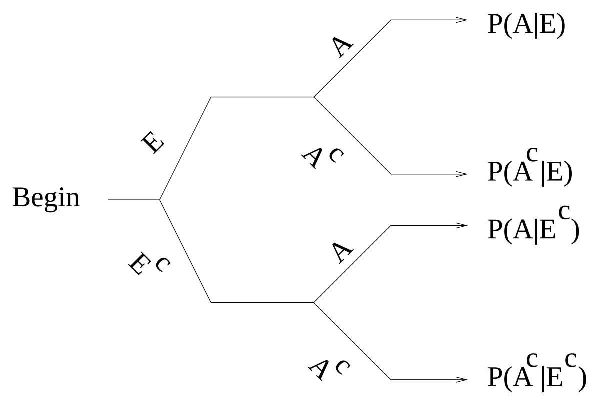
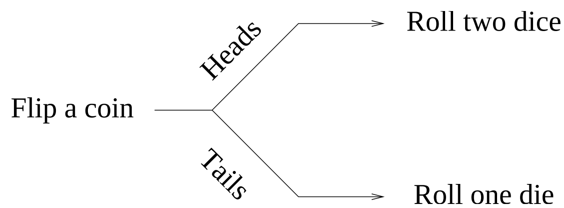
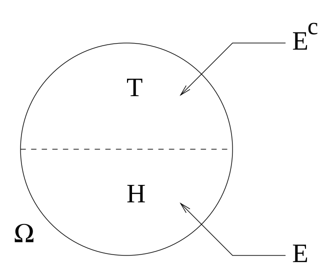


Figure 4 gives a graphical depiction of Theorem 5.

Example 11. *Let E be the event of flipping a coin and getting a head; conversely,* E^c *is the event of getting a tail. The experiment is more complicated than that because depending on the outcome of flipping a coin either two dice are rolled or only one die is rolled as illustrated by the following schematic diagram of the experiment.*



The sample space consisting of flipping a coin is:



Let A be the event that the sum of the faces is 5. The outcome of A might happen by casting one die or it might happen by casting two dice depending on the outcome of flipping the coin. Find $P(A)$.

In itself, the event, A, is complicated enough to make the deduction of its probability a challenge. The solution can be easily derived, however, by decomposing A according to the events of casting one die or two dice.

$$\begin{aligned} P(A) &= P(A|E)P(E) + P(A|E^c)P(E^c) \\ &= (\frac{4}{36})(\frac{1}{2}) + (\frac{1}{6})(\frac{1}{2}) = \frac{5}{36} \end{aligned}$$

If a head occurs from the flip, then two dice are cast, hence $P(A|E) = \frac{4}{36}$, but if a tail occurs, then one die is cast, hence $P(A|E^c) = \frac{1}{6}$.

Example 12. *An instructor of political science wishes to predict, without asking his students, the political composition of his class. In other words, he wants to know the probability that a student would claim himself to be a Democrat, for instance. The instructor knows from the class roster that there are 21 men and 28 women enrolled in the class. From a newspaper article, the instructor learned the political affiliation of men and women across the country. According to the newspaper article, the proportion of men and women by political party is shown below in the table on the left, and according to the class roster, the composition of the political science class is shown in the table on the right.*

	Democrat	Republican
Men	.38	.62
Women	.56	.44

		Proportion
Men	21	$\frac{3}{7}$
Women	28	$\frac{4}{7}$
Total	49	

Given that D represents the event that a student is a Democrat, the instructor essentially wants to find $P(D)$. The event, D, is too complicated for making the computation easy. Consequently, the strategy in solving the problem is to decompose the event into simpler pieces.

Let M be the event that a student is a man; equivalently, M^c is the event that a student is a woman. In order to utilize the information which the instructor saw in the newspaper, he must assume that the political sentiments of college students closely reflect those of the general population.

Before beginning the process of decomposing the event, D, the instructor, to be precise, should define the sample space, Ω. Assuming that a student must be either a Democrat or a Republican and nothing else, the sample space will be like the sample space of flipping 49 coins in which heads will stand for D and tails will stand for R. The sample space will consist of 2^{49} outcomes.

According to the newspaper article, the outcomes are not equally likely. It is obvious that a sample space of this size poses a formidable challenge in answering the instructor's curiosity. But conditioning reduces the size of the sample space. Given that a student is a man reduces the sample space from 2^{49} *outcomes to two outcomes, namely,* $D|M$ *and* $R|M$. *Now it is easy to deduce from the left table that* $P(D|M) = .38$. *Similarly, given that a student is a woman reduces the sample space from* 2^{49} *to two elements,* $D|M^c$ *and* $R|M^c$, *so that* $P(D|M^c) = .56$.

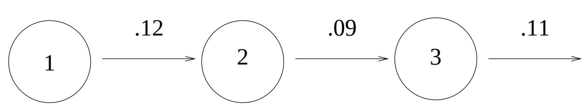
Not all choices of the ancillary event, M, are successful. There is no recipe for inventing a good ancillary event like M. Many times, it is found by trial and error with ingenuity and a little bit of luck. In any case, the utility of conditional probability makes it a very popular technique for incorporating additional information into a problem, so that among other purposes it can be used to divide a difficult event into manageable pieces.

The instructor may proceed in answering his question by doing the computations.

$$\begin{aligned} P(D) &= P(D|M)P(M) + P(D|M^c)P(M^c) \\ &= .38\left(\frac{3}{7}\right) + .56\left(\frac{4}{7}\right) = .482857 \end{aligned}$$

Independence is ideal. To have independent events certainly makes computations of probability much easier. Of course, in the real world, everything is dependent on everything else to some extent. Authors of science fiction like to exploit this when they write about time travel, for instance. In the story, a traveler in time whether the hero or villain might deliberately perturb an event which occurred in the past in a seemly insignificant way. No matter how insignificant it might first appear, the perturbation offers for the delightful author innumerable consequences which when magnified by the author's imagination alter the course of civilization in unexpected ways.

Dependencies might be so slim that for practical purposes, the events may be deemed to be independent. A good example is the successful operation of a system of three components connected in series.



Component 1 fails with probability .12; component 2 fails with probability .09; component 3 fails with probability .11. If any one component fails, then the system will fail. It is assumed that each component is independent of the others. The question which a reliability engineer might ask is: what is the probability that the system will fail?

The system as described by the schematic diagram is an abstract one. The components could represent electrical switches or shipments in a "just-in-time" inventory for a manufacturer or examinations of a particularly stringent process for receiving a professional certification.

Let A be the event that component 1 fails, then $P(A) = .12$. The probability that component 1 functions correctly would then be $P(A^c) = .88$. Let B be the event that component 2 fails, then $P(B) = .09$; $P(B^c) = .91$. And let C be the event that component 3 fails, then $P(C) = .11$; $P(C^c) = .89$.

To find the probability that the system works, we must recognize that all individual components must work, that is, we need to find $P(A^c \cap B^c \cap C^c)$. Now it is crucial to assume that the components work independently. Recall from Theorem 4 that for independence, $P(A \cap B) = P(A)P(B)$, so that

$$P(A^c \cap B^c \cap C^c) = P(A^c)P(B^c)P(C^c) = (.88)(.91)(.89) = .712712$$

What is the probability that the system will fail? We need to find the probability of the complement of the event that the system works. To that end, let S be the event that the system works.

$$P(fails) = 1 - P(works) = 1 - P(S) = 1 - P(A^c \cap B^c \cap C^c) = 1 - .712712 = .287288$$

In a series configuration such as the one discussed above, all three components must work, in order for the system to work. In a parallel configuration, the system will work if any component will work. That is, the system will fail if all components fail. Suppose the three components are connected in parallel rather than in series, then $P(S^c) = P(A^c \cap B^c \cap C^c) = .12(.09).11 = .001188$; therefore, $P(S) = 1 - .001188 = .998812$. A parallel configuration would be used for a system of sensors in a fire extinguisher system, for example. If any sensor is actuated, the system will be turned on.

5 Random Variables

Although defining the sample space is the first step in solving a problem in probability, the sample space usually contains too much information. In the example concerning the tossing of several coins, the sequence of heads and tails might not be of interest; however, the number of heads and the number of tails usually is the only desired information. In the case of tossing three coins, the sample space consists of eight outcomes:

$$\Omega = \{TTT,\ TTH,\ THT,\ THH,\ HTT,\ HTH,\ HHT,\ HHH\}$$

Let E be the event that two and only two heads appear, $E = \{THH,\ HHT,\ HTH\}$. The size of E equals the number of outcomes having two and only two heads. The order of the heads and tails within an outcome is not useful information. If we are only interested in the number of heads, then define $X(\omega)$ to be equal to the number of heads in the outcome, ω. The event, E, can be interpreted to be the set of all outcomes, ω, lying in the sample space, Ω, such that $X(\omega) = 2$. In terms of set notation, $E = \{\omega \in \Omega \textit{ such that } X(\omega) = 2\}$. In this

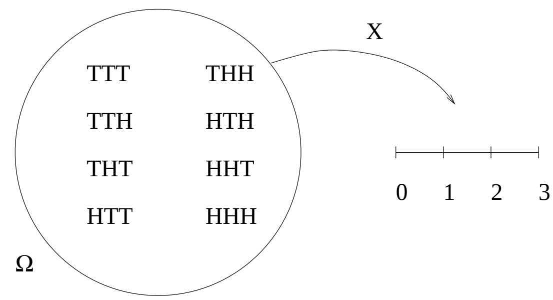


Figure 2

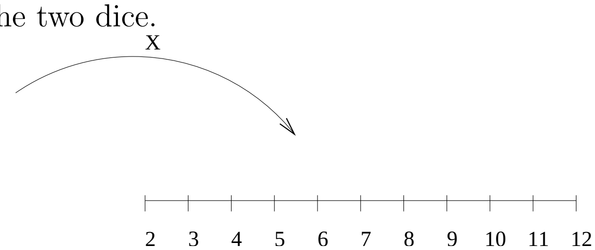
formulation of the event, E, the symbol for *is an element of* appears as $\in$. It is derived from the Greek verb, ἐστί, which means the same as the English word *is*. This notation is too verbose. We will abbreviate *such that* by the common abbreviation, *s.t.*, so that $E = \{\omega \in \Omega \ s.t. \ X(\omega) = 2\}$. Even that abbreviation is still too verbose for many mathematicians who prefer to abbreviate *such that* by the symbol, $|$, as a result, a completely rigorous definition of the event E is: $E = \{\omega \in \Omega | X(\omega) = 2\}$. This is translated into English as: *the event, E, is the set of all outcomes in the sample space, Ω, such that an outcome has exactly two heads*. Whereas the mathematical statements uses fourteen symbols, the English translation uses 86 symbols (including commas). The virtue of mathematics is its parsimony of symbols. As the complexity of events increases, the English becomes proportionally more obscure. A difficult challenge of mathematical comprehension faces the historians of mathematics whenever they read mathematical treatises written as late as the 18^{th} century which usually are composed in Latin with hardly a mathematical symbol. The mathematical works written in English of that era are equally as incomprehensible. The invention and logical application of mathematical symbols which clears away the convoluted prose truly deserves our gratitude.

The probability of the event of getting two heads can be written succinctly as: $P(E) = P(\{\omega \in \Omega | X(\omega) = 2\})$. The use of set notation emphasizes the fact that an event is a set of outcomes of a sample space. In rigorous theoretical developments of probability, set notation is employed to make the discussion absolutely clear and precise to the reader. For the right audience, seasoned mathematicians will take the lazy way by omitting some symbols. It is not uncommon to see the above statement written as: $P(E) = P(X = 2)$. Whichever notation is used, it must bring to mind the connotations of sets.

With X defined to count the number of heads in an outcome, for example $X(\{TTT\}) = 0$ and $X(\{TTH\}) = 1$, there are eight elements in the sample space, but there are only 4 values of X, namely 0, 1, 2, and 3. In effect, X reduces the size of the sample space as depicted in Figure 2 and brings about a simplification in the name of clarity but at the expense of loosing unwanted information like the actual sequence of the T's and H's.

In another illustration of the resolution of a sample space into something simpler through the use of a random variable, an experiment of rolling two fair dice is performed. Let X equal the sum of the faces of the two dice.

$$\Omega = \begin{Bmatrix} 11 & 12 & 13 & 14 & 15 & 16 \\ 21 & 22 & 23 & 24 & 25 & 26 \\ 31 & 32 & 33 & 34 & 35 & 36 \\ 41 & 42 & 43 & 44 & 45 & 46 \\ 51 & 52 & 53 & 54 & 55 & 56 \\ 61 & 62 & 63 & 64 & 65 & 66 \end{Bmatrix}$$



The possible outcomes of X are: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12. Rather than 36 elements, we deal with eleven values of X. In terms of a random variable, let E be the event such that X=5, that is: $E = \left\{14,\ 23,\ 32,\ 41\right\} = \{\omega \in \Omega | X(\omega) = 5\}$, so that

$$P(E) = P(\{\omega \in \Omega | X(\omega) = 5\}) = \frac{size\ of\ E}{size\ of\ \Omega} = \frac{4}{36} = \frac{1}{9}$$

This set function, $X(\omega)$, with which to consolidate useful information of a sample space is recognized by a name.

Definition 11. *A function, X, which maps an outcome of the sample space to a number on the real line is called a **random variable**.*

Example 13. *Toss three fair coins. The sample space is the usual:*

$$\Omega = \{TTT,\ TTH,\ THT,\ THH,\ HTT,\ HTH,\ HHT,\ HHH\}$$

Let X be the number of heads and Y be the number of tails in an outcome.

1. $E_1 = \{\omega \in \Omega | X(\omega) = 1\} = \{HTT,\ THT,\ HTT\}$, *so that* $P(E_1) = P(X = 1) = \frac{3}{8}$.
2. $E_2 = \{\omega \in \Omega | X(\omega) > 2\} = \{HHH\} \to P(X > 2) = \frac{1}{8}$.
3. $E_3 = \{\omega \in \Omega | 1 \leqslant Y(\omega) \leqslant 2\} = \{TTH,\ THT,\ HTT,\ THH,\ HTH,\ HHT\} \to P(1 \leqslant Y \leqslant 2) = \frac{6}{8} = \frac{3}{4}$.
4. $P(\{\omega \in \Omega | 3X(\omega) + 1 = 7\}) = P(\{\omega \in \Omega | 3X(\omega) = 6\}) = P(\{\omega \in \Omega | X((\omega)) = 2\}) = P(\{THH,\ HTH,\ HHT\}) = \frac{3}{8}$. *In an abbreviated and equivalent way:* $P(3X + 1 = 7) = P(X = 2) = \frac{3}{8}$.
5. $E_5 = \{\omega \in \Omega | Y(\omega)^2 = 4\} = \{\omega \in \Omega | Y(\omega) = 2\} = \{TTH,\ THT,\ HTT\}$, *so that,* $P(\{\omega \in \Omega | Y(\omega)^2 = 4\}) = \frac{3}{8}$. *An abbreviated version of the expression is:* $P(Y^2 = 4) = P(Y = 2) = \frac{3}{8}$.

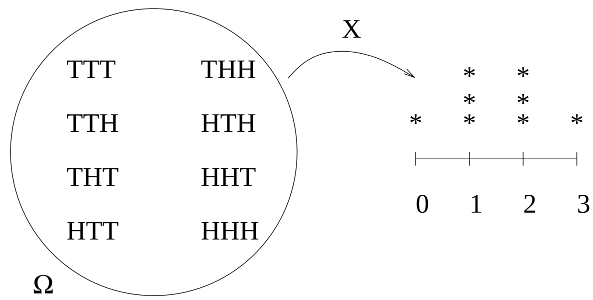


Figure 3

Through the use of random variables, events can be viewed from a different perspective in that they liberate us from the limitations of a natural language like English and make the vast collection of mathematical tools relating to algebra and calculus ready for use. As a result, the explanation of a phenomenon is only limited by our mathematical ingenuity. Random variables, for example, bring meaning to quantum mechanics and statistical mechanics; they allow electrical engineers a way to study the source of noise and to remove it from an electronic device; they allow economists to formulate models of the economy. By associating data with random variables, a set of data can be firmly grasped; an analyst can invoke the enormous store of mathematics and manipulate a set of data in ways to probe its origins which could never be imagined possible with plain English.

Let us look at this experiment of flipping three fair coins as if we were looking at statistical data. The random variable, X, maps each outcome to a number. There may be more than one outcome which is mapped to the same number like the outcomes $\{TTH\}$, $\{THT\}$, and $\{THT\}$ which are each mapped to 1. By recording each mapping above each value of X as is done in Figure 3, a picture which looks like a histogram emerges. The set of values of X from mapping all the outcomes of the sample space is $\left\{0 \;\; 1 \;\; 1 \;\; 1 \;\; 2 \;\; 2 \;\; 2 \;\; 3\right\}$. It looks like a set of data; it looks like something which we should analyze; it beckons us to calculate its mean, variance, and median. We, therefore, find that

$$\mu = \frac{0+1+1+1+2+2+2+3}{8} = \frac{12}{8} = \frac{3}{2}$$

$$\begin{aligned}\sigma^2 &= \frac{(0-\frac{3}{2})^2 + (1-\frac{3}{2})^2 + (1-\frac{3}{2})^2 + (1-\frac{3}{2})^2 + (2-\frac{3}{2})^2 + (2-\frac{3}{2})^2 + (2-\frac{3}{2})^2 + (3-\frac{3}{2})^2}{8} \\ &= \frac{3}{4}\end{aligned}$$

and the median is $\frac{3}{2}$.

The mean and median are not necessarily values of X. They are numbers that reflect the center of mass of the possible values of X.

Let X count the number of heads occurring in the outcomes of flipping three fair coins. It was deduced already that:

$P(X = 0) = \frac{1}{8}$
$P(X = 1) = \frac{3}{8}$

$P(X = 2) = \frac{3}{8}$

$P(X = 3) = \frac{1}{8}$

These fractions resemble the fractions which were used to calculate the mean, μ, of the set of values of X. Pursuing this observation further, an algebraic manipulation of $\mu = \frac{0+1+1+1+2+2+2+3}{8} = \frac{3}{2}$ transforms this expression of μ into one involving probabilities. That is,

$$\mu = \frac{1(0) + 3(1) + 3(2) + 1(3)}{8} = 0(\frac{1}{8}) + 1(\frac{3}{8}) + 2(\frac{3}{8}) + 3(\frac{1}{8})$$

then substituting the fractions with the probabilities which are displayed above, μ can be written as

$$\mu = 0P(X = 0) + 1P(X = 1) + 2P(X = 2) + 3P(X = 3)$$

The random variable, X, produces a set of values which in a certain sense resembles data from which a special mean denoted by $E[X]$ can be obtained. Instead of using the term, *mean*, for the center of mass of the distribution of values of a random variable as is done for statistical data, the phrase, *expected value*, is the term used for the center of mass of a random variable in the field of probability.

Definition 12. *The* ***expected value*** *of a random variable, X, is defined to be:*

$$E[X] = \sum_{\substack{all\ possible\\ values\ of\ X}} kP(X = k).$$

(The summation $\sum$ is taken over all possible values of X)

The mean and expected value convey the same connotation. The mean is the center of mass of a set of data; the expected value is the center of mass of a random variable.

Let X be the number of heads in an outcome of flipping three fair coins, then:

$$E[X] = 0P(X = 0)+1P(X = 1)+2P(X = 2)+3P(X = 3) = 0(\frac{1}{8})+1(\frac{3}{8})+2(\frac{3}{8})+3(\frac{1}{8}) = \frac{3}{2}$$

In the same vein of developing the notion of expected value along the lines of finding the mean of a set of data, the variance of a random variable imitates the definition of the population variance.

Definition 13. *The* ***variance*** *of a random variable, X, is defined to be:*

$$var(X) = \sum_{\substack{all\ possible \\ values\ of\ X}} (k - E[X])^2 P(X = k).$$

(The summation is taken over all possible values of X)

Let X be the number of heads in an outcome of flipping three fair coins. The variance of X is:

$$\begin{aligned} var(X) &= (0 - \frac{3}{2})^2 P(X = 0) + (1 - \frac{3}{2})^2 P(X = 1) + (2 - \frac{3}{2})^2 P(X = 2) + (3 - \frac{3}{2})^2 P(X = 3) \\ &= (\frac{9}{4})(\frac{1}{8}) + (\frac{1}{4})(\frac{3}{8}) + (\frac{1}{4})(\frac{3}{8}) + (\frac{9}{4})(\frac{1}{8}) \\ &= \frac{24}{32} = \frac{3}{4} \end{aligned}$$

A probability is defined by some specific characteristic of the phenomenon. In the case of flipping a fair coin, the fundamental characteristic of the process is the property that the coin is fair. If the coin is loaded, then a different probability is induced. In either case, a random variable may have many different probabilities associated with it depending on the nature of the phenomena. For a given phenomenon and for each value of X, there is a probability. That complete collection of probabilities which are associated with the random variable and induced by a specific phenomenon is given a special name.

Definition 14. *The set of values* $\{P(X = 0),\ P(X = 1),\ \ldots,\ P(X = n)\}$ *is called the* ***probability distribution*** *or the* ***probability mass function*** *of the random variable X.*

The great utility of a random variable lies in facilitating the precise description of an event. The size of an event relative to the size of the sample space is a probability. Each association of a random variable with a phenomenon induces a probability distribution; hence when a random variable describes an event, it comes with a probability distribution. The axioms of probability apply as well to random variables as to events. According to the second axiom of probability, $P(\Omega) = 1$. With that observation, the following theorem is an immediate consequence.

Theorem 6.

$$\sum_{\substack{all\ possible \\ values\ of\ X}} P(X = k) = 1$$

Proof. The events $\{\omega \in \Omega | X(\omega) = i\}$ and $\{\omega \in \Omega | X(\omega) = j\}$ are disjoint. For example, it is logically impossible for an outcome to have exactly one head and exactly two heads at the same time. The meaning of the statement to be proved becomes clearer when it is formally written in set notation. That is,

$$P(X=0)+P(X=1)+P(X=2)+\ \ldots\ +P(X=n)=$$
$$P(\{\omega\in\Omega|X(\omega)=0\})+P(\{\omega\in\Omega|X(\omega)=1\})+\ \ldots\ +P(\{\omega\in\Omega|X(\omega)=n\})$$

Because the events are disjoint, we now invoke Theorem 1, in order to consolidate the sum of probabilities into a probability of a union of disjoint events. Explicitly, the probability of that union of events can be written as:

$$P(\{\omega\in\Omega|X(\omega)=0\}\cup\{\omega\in\Omega|X(\omega)=1\}\cup\ \cdots\ \cup\{\omega\in\Omega|X(\omega)=n\})$$

But the union of all possible events constitutes the sample space, Ω; therefore,

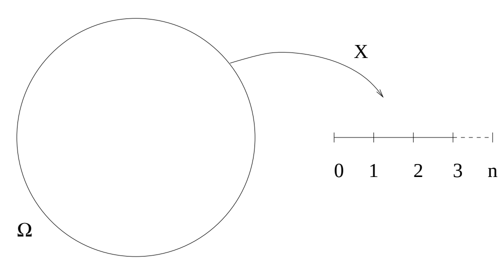
$$\sum_{\substack{all\ possible\\ values\ of\ X}} P(X=k)=P(\Omega)=1$$

■

The sum of a probability distribution is always equal to 1. If the sum is not equal to 1, then the distribution is defective. It is always prudent to verify that a given probability distribution produces a 1 when all of its terms are added together. It is especially prudent to do so when solving a problem or deriving a customized statistical theory to explain a new phenomenon.

There are two basic kinds of random variables: discrete and continuous.

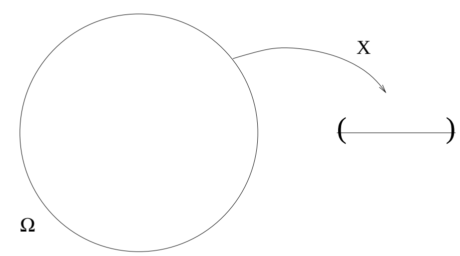
Definition 15. *If the random variable, X,*



maps elements of Ω to a finite set of numbers, then X is called a ***discrete random variable****.*

On the other hand, a random variable might have an infinite number of possible values and that greatly complicates matters as the next definition suggests.

Definition 16. *If the random variable, X,*



*maps elements of Ω to a set of intervals, then X is called a **continuous random variable**.*

There are special random variables which, though they have an infinitely denumerable range like the Geometric distribution, are nonetheless classified as discrete random variables and appear to defy the meaning of discrete random variables. Very strange paradoxes arise in any discussion of infinities. There are theorems which seem unbelievable such as the proposition that there are an infinite number of infinities. In and among themselves, the properties of infinity form a fascinating subject for study.

It is easier to define the meaning of a continuous random variable by giving specific examples like temperature, length, time, weight, and voltage. These are attributes of things belonging to the physical universe. Paradoxically, the universe is composed of atoms and photons which are discrete entities about which energy and distance between and about them are quantized. Strictly speaking then, there are no continuous random variables in reality; they exist only in our imagination. Should we continue along this line of reasoning, we will insensibly drift into a deep philosophical discussion. There are many intriguing philosophical questions rooted in the natural sciences which are worth pondering, but it will be better for us to circumvent this sort of problematic discussion and to follow our innate intuition about the meaning of continuity and infinite things and let the physicists and logicians resolve such inscrutable contradictions. By leaving that subject behind, we are free to discuss the properties of probability distributions.

Besides the properties listed below which pertain to the discrete random variable, there are analogous properties for distributions of continuous random variables as we will later see.

Observations 1.

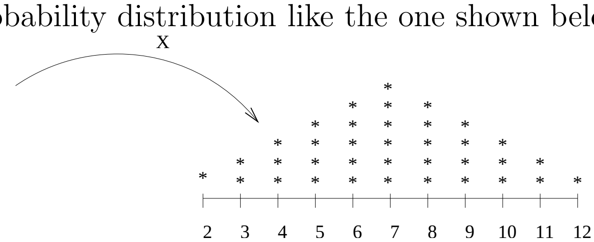
1. $P(X = i) \geqslant 0$.
2. $\displaystyle\sum_{\substack{\text{all possible} \\ \text{values of } X}} P(X = k) = 1$.
3. $P(X \leqslant k) = P(X = 0) + P(X = 1) + \ \ldots \ + P(X = k - 1) + P(X = k)$.
4. $P(X > k) = 1 - P(X \leqslant k)$.
5. $P(X = k) = P(X \leqslant k) - P(X \leqslant k - 1)$.

Besides making a complete listing of the probabilities which are associated with a random variable, in order to define a probability distribution, there is another and equivalent way to characterize a probability distribution. It is done by means of listing the partial sums of the probability mass function. These partial sums are known as the cumulative distribution function.

Definition 17. *The* ***cumulative distribution function****, F(c), is defined to be:* $P(X \leqslant c) = F(c) = \sum_{k=0}^{c} P(X = k)$*, and it is abbreviated by cdf.*

As was noted earlier, a random variable is defined on a sample space, and associated with it there is a probability distribution perhaps among many others. A picture of a probability distribution looks like a histogram. For example, when X is the sum of the faces of two fair dice, it induces a probability distribution like the one shown below.

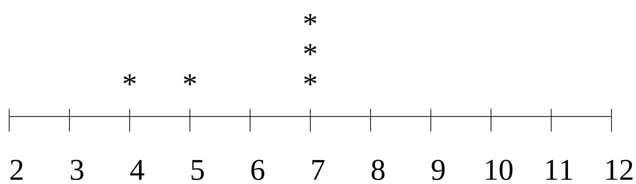
$$\Omega = \left\{ \begin{array}{cccccc} 11 & 12 & 13 & 14 & 15 & 16 \\ 21 & 22 & 23 & 24 & 25 & 26 \\ 31 & 32 & 33 & 34 & 35 & 36 \\ 41 & 42 & 43 & 44 & 45 & 46 \\ 51 & 52 & 53 & 54 & 55 & 56 \\ 61 & 62 & 63 & 64 & 65 & 66 \end{array} \right\}$$



What looks like a histogram is actually a probability distribution. Care must be observed to distinguish a histogram and a probability distribution. A histogram of data is to statistics what a probability distribution of a random variable is to probability. If a probability distribution resembles a histogram, then we have established the desired bridge between statistics and probability. That connection is the link which will allow us to work backwards from the data to explain a characteristic of the population in a process called inference. The objective of the statistician is to make a valid connection between a probability distribution and a histogram of the data. The association of a probability distribution with a histogram is an important source of controversy. The association is often fraught with uncertainty and it requires meticulous substantiation otherwise no one will believe that the association is valid. Once, if ever, the association is deemed valid, then the job of the statistician is essentially finished.

Tabulated below are the results of an actual experiment in which two fair dice were tossed five times. In the table, X represents the sum of the faces. A picture of the data accompanies the data.

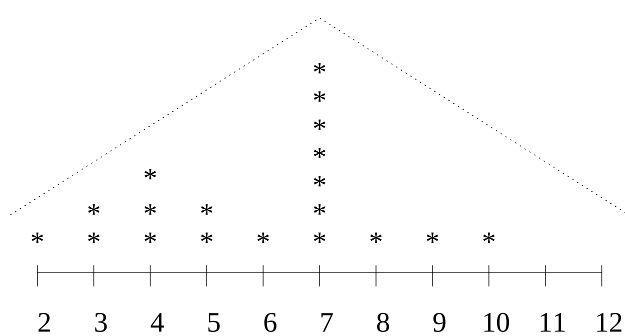
Throw	Outcome	X
1	25	7
2	32	5
3	43	7
4	16	7
5	31	4



In comparison to the picture of the probability distribution which was shown earlier, this histogram of the data remotely resembles the triangular shape of the probability distribution of X. Suppose the dice were thrown 14 more times. The tabulation of all 19 casts is shown here in Table 1:

Table 1

Throw	Outcome	X	Throw	Outcome	X	Throw	Outcome	X
1	25	7	8	25	7	14	61	7
2	32	5	9	13	4	15	52	7
3	43	7	10	13	4	16	12	3
4	16	7	11	32	5	17	35	8
5	31	4	12	52	7	18	63	9
6	64	10	13	11	2	19	21	3
7	15	6						



With additional points inserted into the histogram, the histogram begins to resemble the triangular shape of the probability distribution. The histogram will never co-incide with a probability distribution because one is derived from experimental data while the other is an ideal. A natural question suggests itself by this illustration, namely, when should the process of throwing the dice stop before the resemblance of the histogram with the probability distribution is sufficiently convincing? The answer to that question is difficult in practice to find, but a flavor of it will be given in the chapter on testing the hypothesis for the goodness-of-fit.

6 Common Distributions

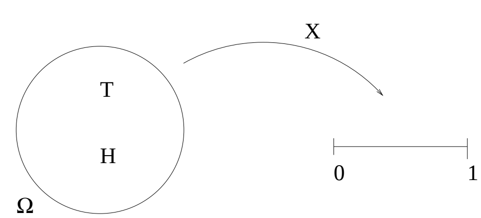
Discrete Random Variable	Continuous Random Variable
Bernoulli	Uniform
Uniform	Normal
Binomial	Student's t
Multinomial	X^2
Triangle	F
Geometric	Gamma
Hypergeometric	Exponential
Poisson	Beta

Associated with every random variable, there is a probability distribution. Some probability distributions occur so often that they are given names. A list of the common distributions is shown above.

7 Discrete Random Variables

7.1 Bernoulli

Only two outcomes characterize the Bernoulli distribution. The correct pronunciation of *Bernoulli* puts the accent on the last syllable: Ber·noul·lí.



The sample space consists of two outcomes: pass-fail, head-tail, success-failure, on-off, 0-1. Let p be the probability of success, so that 1-p is the probability of failure. If X counts the number of successes, then the possible values of X are 0 and 1. The event corresponding to $\{\omega \in \Omega | X(\omega) = 1\}$ is the event of getting a head, so that P(X=1)=p. Similarly, the probability of getting a tail is P(X=0)=1-p=q. It is a common convention to denote 1-p by q.

The formulas which we are about to derive for the expected value and the variance of a Bernoulli random variable are specific only to the Bernoulli distribution and to no other. The derivation begins in the same manner as with the derivation of expected value and variance for any other distribution by beginning with the basic definitions of expected value and variance. From the definition of expected value we may write:

$$E[X] = \sum_{k=0}^{1} kP(X = k) = 0P(X = 0) + 1P(X = 1) = p$$

and similarly

$$\begin{aligned} var(X) &= \sum_{k=0}^{1} (k - E[X])^2 P(X = k) = (0 - p)^2 P(X = 0) + (1 - p)^2 P(X = 1) \\ &= p^2(1 - p) + (1 - p)^2 p = p(1 - p)(p + 1 - p) = p(1 - p) \\ &= pq \end{aligned}$$

There is no need anymore to calculate $E[X]$ and $var(X)$ for a Bernoulli random variable. It is sufficient to recognize that the random variable under consideration is associated with two and only two outcomes then one may use the formulas shown below for computing the expected value and the variance.

Bernoulli b(1,p)
$E[X] = p$
$var(X) = pq$

Example 14. *A question on a multiple choice examination has five possible answers, one of which is correct. What is the probability that a student who merely guesses will choose the correct answer? ANS:* $\Omega = \{C,\ I\}$ *where C is correct and I is incorrect.*

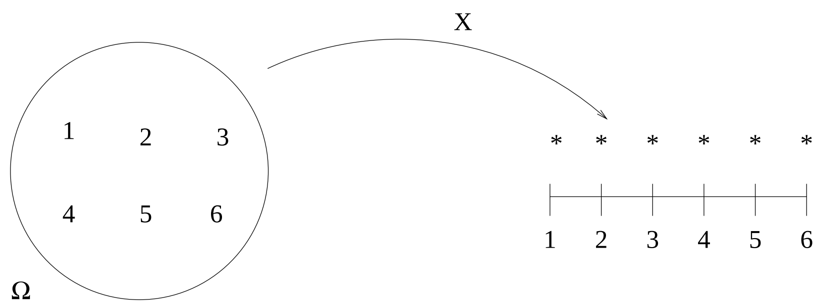
Let $X(\omega) = \begin{cases} 1 & \text{if } \omega \text{ is correct} \\ 0 & \text{otherwise} \end{cases}$

The assumption that the student guesses implies equally likely outcomes as in drawing a white ball out of an urn of five balls, one of which is white and the others red; therefore, the probability of success is $P(\{\omega \in \Omega | X(\omega) = 1\}) = \frac{1}{5}$.

Find $E[X]$ and $var(X)$. ANS: X is a Bernoulli random variable; therefore, $E[X] = p = \frac{1}{5}$ and $var(X) = pq = \frac{1}{5}\frac{4}{5} = \frac{4}{25}$.

7.2 Uniform

The Uniform distribution is characterized by the fact that for all values of the random variable, X, the probabilities are the same. A precise description is the following. A random variable, X, is distributed as a discrete Uniform distribution if $P(X = k) = \frac{1}{n}$ for every value of X. For example, the sample space of tossing one fair die consists of six outcomes. If the random variable, X, gives the number on the face of the die, then the sample space of six outcomes is mapped to six possible values of X as portrayed in the accompanying figure.



Having stated that the die is fair, the probability distribution of X is simply,

$$P(X = 1) = P(X = 2) = P(X = 3) = P(X = 4) = P(X = 5) = P(X = 6) = \frac{1}{6}$$

.

The discrete Uniform distribution occurs in so many varieties that attempts to derive a general formula for the expected value and the variance are not too helpful. In the special case, however, in which the possible values of X are consecutive from 1 to n then, $P(X = k) = \frac{1}{n}$ for $k = 1,\ 2,\ 3,\ \ldots,\ n$. By the definition of expected value, $E[X] = \sum_{k=1}^{n} kP(X = k) = 1(\frac{1}{n}) + 2(\frac{1}{n}) + \ \ldots\ + n(\frac{1}{n}) = \frac{1+2+\ \ldots\ +n}{n} = \frac{n+1}{2}$.

The derivation of the formula for the variance in this special case of consecutive values of X begins as usual with the definition of variance.

$$\begin{aligned} var(X) &= \sum_{k=1}^{n}(k - \frac{n+1}{2})^2 P(X = k) = \left(\sum k^2 - 2\frac{(n+1)}{2}\sum k + \frac{(n+1)^2}{4}\right)\frac{1}{n} \\ &= \frac{\sum k^2}{n} - \frac{(n+1)^2}{4} = \frac{n(n+1)(2n+1)}{n6} - \frac{(n+1)^2}{4} \\ &= \frac{(n-1)(n+1)}{12} \end{aligned}$$

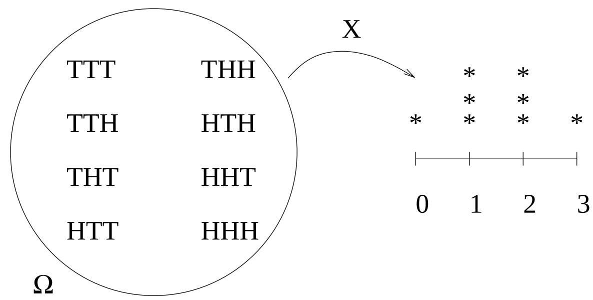
Example 15. *In casting one fair die, n=6, and the possible values of X are consecutive from 1 to 6.* $E[X] = \frac{n+1}{2} = \frac{7}{2}$ *and* $var(X) = \frac{(n-1)(n+1)}{12} = \frac{35}{12}$*. If the values of X had not been consecutive, then it would have been necessary to resort to the basic definitions of expected value and variance to calculate them.*

Usually, one resorts to the definitions of expected value and variance to find the numerical values for the Uniform distribution but in a slightly more generalized version of the special case discussed above for $k = 1, 2, 3, \ldots, n$ is the one where k starts at a and sequentially goes to b. For this Discrete Uniform distribution, $P(X = k) = \frac{1}{b-a+1}$, $E[X] = \frac{a+b}{2}$ and $var(X) = \frac{(b-a)(b-a+2)}{12}$. If the parameter, a, were set to 1 and b set to n, then these last formulas collapse to the ones cited above.

Discrete Uniform (a,b)
$E[X] = \frac{a+b}{2}$ $var(X) = \frac{(b-a)(b-a+2)}{12}$

7.3 Binomial

Flipping three coins generates eight possible outcomes. A flip of a coin is referred to as a trial. This experiment of flipping three coins involves three trials for which each trial has two possibilities: either a head or a tail. Three trials, each providing two possibilities, produce a total of $2^3 = 8$ outcomes to constitute the sample space. In general, if an experiment involves n independent trials for which each trial has two possible outcomes, then a random variable defined on this resulting sample space is called a Binomial random variable. The Binomial distribution is characterized by n independent trials for which each trial has two outcomes. A schematic diagram of the sample space being mapped by X and the induced probability distribution for a fair process appears below.



The probability that a Binomial random variable equals k is given by formula (1). This formula is applicable to any Binomial random variable which is associated with n trials such that the probability of obtaining a success on any trial is p.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for} \quad k = 0,\ 1,\ 2,\ \ldots,\ n \tag{1}$$

A random variable which follows a Binomial distribution is denoted by $X \sim b(n, p)$. The symbol , $\sim$, is translated in English as "*is distributed as*". A random variable, X, which is distributed as a Binomial distribution with 3 trials and a probability of success, $\frac{1}{2}$, would be written as: $X \sim b(3, \frac{1}{2})$.

The Binomial and Bernoulli distributions are related. In fact, the Bernoulli distribution is a special case of the Binomial distribution. A Bernoulli random variable, X, is characterized by having one trial with a probability of success, p. In other words, $X \sim b(1, p)$. From another perspective, the Binomial random variable, $X \sim b(n, p)$, is the composition of a sum of n independent Bernoulli random variables. The Bernoulli random variable is the simplest of the discrete random variables, and individual elements of a collection of Bernoulli random variables are often used in clever ways to build more complicated distributions like the Binomial distribution. If there exists such a distribution like the Binomial distribution, then there ought to be something like a Trinomial distribution which would be characterized by n trials with three outcomes per trial, or a Quadnomial distribution for which there would be four outcomes per trial, or a Quintnomial distribution with five outcomes per trial, and so on. All of these generalized versions of the Binomial distribution collectively come under the name, Multinomial distribution.

7.4 Mathematical Interlude on Counting

Suppose five coins are tossed. One outcome might be THTTH. Let X be that random variable which counts the number of heads in an outcome, then X({THTTH})=2. But there are other outcomes having two heads like HHTTT. A natural question to ask is: how many outcomes have two heads? By enumerating all possible outcomes which have two heads, the answer is 10. After performing many such enumerations, a pattern emerges which suggests a rule for counting all combinations of k heads from n coins. Under this rule, the enumeration is easy and transparent.

The science of counting is called combinatorics. It is a branch of mathematics which is surprisingly challenging but which produces remarkably useful results. A basic unit in the study of combinatorics is the permutation. A permutation of a list of letters is simply a rearrangement of them. For example, the three letters, abc, can be permuted in six ways as illustrated below:

$$abc \to acb \to cab \to cba \to bca \to bac$$

The permutations of these three letters occur $3 \times 2 \times 1 = 6$ ways. The first slot in the list may be filled with any one of three letters, but once a letter is picked to fill the first slot, there are only two candidates left to fill the second slot. When both the first and second slots are fill, there is only one possibility left to fill the third slot. With that reasoning, the number of permutations can be calculated. For example, four letters can be permuted in $4 \times 3 \times 2 \times 1 = 24$ ways. The multiplication of a consecutive sequence of numbers in descending order occurs so often in combinatorics that it is given a name.

Definition 18. *$n! = n(n-1)(n-2)\cdots(2)(1)$. n! is called n **factorial**.*

Example 16.

1. $3! = 3 \times 2 \times 1 = 6$
2. $4! = 4 \times 3 \times 2 \times 1 = 24$
3. $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$
4. $52! = 8 \times 10^{67}$

In the last example, 52! is the number of permutations of a list of 52 letters or 52 cards in a standard poker deck. It is a simple task by means of a hand calculator to calculate that $52! = 8 \times 10^{67}$. 52! is a huge number. The image of a string of pearls will demonstrate what huge means in this context. A substitution of the smallest atom for each pearl while leaving no space between them may not produce a necklace as attractive as a string of pearls, but this string of 52! hydrogen atoms will more than encircle the known universe.

In the process of permuting objects, the order of the objects must be considered, but, when the order is irrelevant, the number of combinations of choosing k objects from a total of n objects is $\binom{n}{k}$.

Definition 19. *Let $\binom{n}{k}$ denote the number of subsets of size k that can be made from a set of size n. We say, "n choose k" for $\binom{n}{k}$.*

Example 17. *Let $A = \{a,\ b,\ c,\ d\}$*

1. *The number of subsets of size 1 which can be created from A is 4. These four subsets are:* $\{a\}$, $\{b\}$, $\{c\}$, $\{d\}$; *therefore,* $\binom{4}{1} = 4$.

2. *The only subsets of A which consist of two elements are:*

$$\{ab\},\ \{ac\},\ \{ad\},\ \{bc\},\ \{bd\},\ \{cd\}$$

therefore, $\binom{4}{2} = 6$.

3. *These four subsets,* $\{bcd\}$, $\{acd\}$, $\{abd\}$, $\{abc\}$ *imply that* $\binom{4}{3} = 4$ *which should also be equal to* $\binom{4}{1}$ *because the subsets,* $\{bcd\}$, $\{acd\}$, $\{abd\}$, $\{abc\}$ *are the complements of* $\{a\}$, $\{b\}$, $\{c\}$, $\{d\}$.

4. *The set, A, is a subset of itself; therefore,* $\binom{4}{4} = 1$.

5. *What is* $\binom{4}{0}$*? The empty set,* $\{\} = \emptyset$, *is a subset of every set including the set, A; therefore* $\binom{4}{0} = 1$.

From the experience of working with enough examples of counting combinations in terms of permuting objects, a simple formula emerges.

Theorem 7. $\binom{n}{k} = \frac{n!}{(n-k)!k!}$

A simple device know as Pascal's Triangle is useful for calculating small values of $\binom{n}{k}$. Rows correspond to n and the position within a row corresponds to k. The rows and positions are counted starting with 0. For example, $\binom{4}{2} = 6$.

Pascal's Triangle

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1

The binomial coefficient, $\binom{n}{k}$, plays a prominent role in the formula for calculating the probabilities of a Binomial distribution. It is a symbol that occurs frequently in mathematics, and its origins lie in the expansion of the binomial, x+y, raised to the n^{th} power. The binomial formula comes from elementary algebra and is written as: $(x+y)^n = \sum_{i=0}^{n} \binom{n}{k} x^i y^{n-i}$. If x and y are both set to 1, then $2^n = \sum_{i=0}^{n} \binom{n}{k}$, but $\sum_{i=0}^{n} \binom{n}{k}$ is the sum of the number of all subsets of size k taken from a set of size n. From the binomial formula, therefore, it is deduced that the total number of possible subsets, in other words, the size of the power set is 2^n.

Another mathematician, however, might take a different approach to arrive at the same conclusion. He might argue that we know from our discussions of flipping n distinguishable coins that the sample space consists of 2^n outcomes. There are two possibilities that may be assigned to each of the n slots in an outcome: either a head or a tail. Hence there are $2 \times 2 \times \cdots \times 2 = 2^n$ possible outcomes. Each coin in a string of n coins occurs either with a head or by a tail. Each string of heads and tails correspond to a subset of the sample space, Ω. With this reasoning, the sample space of flipping n coins can be interpreted as a collection of listings of all possible subsets of a set of n things. In other words, the sample space which has a size of 2^n is in one-to-one correspondence with the power set of n objects; therefore, the size of the power set must be 2^n. But the size of the power set is $\sum_{i=0}^{n} \binom{n}{k}$; therefore, $\sum_{i=0}^{n} \binom{n}{k} = 2^n$.

Two different arguments produce the same equation. The former argument may be called the *analytical argument* whereas the second argument could be called the intuitive or *abstract argument*. Both arrive at the same conclusion. Sometimes the analytical approach is the only feasible approach in solving a problem; sometimes the abstract approach is the only feasible approach. Many times both are feasible with the consequence that some remarkable equivalences involving very complicated formulas are produced. One formula will come from an analytical demonstration and another formula will come from abstract reasoning. When both arguments produce the same valid conclusion, the formulas must be equal. This use of playing one argument against another gives a flavor of the kind of clever techniques which are used in the study of combinatorics.

7.5 Return to Binomial Distribution

If $X \sim b(n, p)$, then it has an expected value and a variance. Its expected value is derived directly from the definition: $E[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = np$. This simple answer is all the more remarkable when viewed in the light of the lengthy algebraic manipulations which will not be shown but are required to produce that result. A much greater mathematical challenge is presented by the derivation of the variance. A direct algebraic reduction of the formula for the variance from the basic definition of the variance is usually too challenging to present in an introductory course of probability until the topic of moment generating functions has been adequately discussed. At that time, the formula for the variance with the aid of calculus is easily derived and is given by $var(X) = \sum_{k=0}^{n}(k-np)^2\binom{n}{k}p^k(1-p)^{n-k} = np(1-p) = npq$.

Binomial b(n,p)
$E[X] = np$ $var(X) = npq$

8 Computations Using Probability Distributions

Computations for discrete random variables rely on the properties listed on page 27.

Example 18. *Let X be a Bernoulli random variable such that X=1 with p=.2 and X=0 with 1-p=.8; that is* $X \sim b(1, .2)$.

Without looking at the definition of expected value but instead using the formula of expected value, $E[X] = p = .2$. *Likewise, from the formula for the variance of a Bernoulli random variable,* $var(X) = pq = p(1-p) = .16$ *and the standard deviation is* $std = \sqrt{.16} = .4$.

Example 19. *Let Y=1, 2, 3, 4, 5 be a random variable that follows a Uniform distribution. Since Y is distributed as a Uniform distribution, then* $P(Y = k) = \frac{1}{5}$ *for k=0, 1, 2, 3, 4, and 5.*

1. $P(Y \leqslant 2) = P(Y = 1) + P(Y = 2) = \frac{1}{5} + \frac{1}{5} = \frac{2}{5}$.
2. *Resorting to the definition of expected value:* $E[Y] = 1(\frac{1}{5})+2(\frac{1}{5})+3(\frac{1}{5})+4(\frac{1}{5})+5(\frac{1}{5}) = 3$. *Or by recognizing that within our midst there is the Discrete Uniform random variable having consecutive values from 1 to 5, so that* $E[Y] = \frac{1+5}{2} = 3$.
3. *For the variance:* $var(Y) = (1-3)^2\frac{1}{5}+(2-3)^2\frac{1}{5}+(3-3)^2\frac{1}{5}+(4-3)^2\frac{1}{5}+(5-3)^2\frac{1}{5} = 2$. *Or* $var(Y) = \frac{(5-1)(5-1+2)}{12} = 2$.

Example 20. *Let* $Z \sim b(7, .6)$. *From the formula for the probability of a Binomial* *distribution:*

$$\begin{aligned}
P(Z=0) &= \binom{7}{0}.6^0(.4)^{7-0} = 1(.4)^7 = .00164 \\
&\vdots \\
P(Z=5) &= \binom{7}{5}.6^5(.4)^2 = 21(.07776)(.16) = .26127 \\
P(Z=6) &= \binom{7}{6}.6^6(.4)^1 = 7(.046656)(.4) = .13064 \\
P(Z=7) &= \binom{7}{7}.6^7(.4)^0 = 1(.02799) = .02799
\end{aligned}$$

1. $P(Z > 5) = P(Z = 6) + P(Z = 7) = .13064 + .02799 = .15863.$

2. $P(Z \leqslant 5) = 1 - P(Z > 5) = 1 - .15863 = .84137$. *The direct way produces the same result:*

   $$P(Z \leqslant 5) = P(Z=0)+P(Z=1)+P(Z=2)+P(Z=3)+P(Z=4)+P(Z=5) = .84137$$

3. $P(Z \leqslant 4) = P(Z = 0) + P(Z = 1) + P(Z = 2) + P(Z = 3) + P(Z = 4) = .58010.$

   Let us find $P(Z = 5)$ *in another way. Invoking the property,* $P(X = k) = P(X \leqslant k) - P(X \leqslant k-1)$ *found on page 27,* $P(Z = 5) = P(Z \leqslant 5) - P(Z \leqslant 4) = .84137 - .58010 = .26127$ *which is the same number gotten above.*

4. *The expected value is:* $E[Z] = np = 7(.6) = 4.2.$

5. *And the variance is:* $var(Z) = npq = 7(.6)(.4) = 1.68.$

That important property, $P(X = k) = P(X \leqslant k) - P(X \leqslant k-1)$, will be used to solve the next problem. Although the random variable $Z \sim b(7, .6)$ is rather simple, the Binomial distribution can often lead to formidable computations. Binomial random variables for n up to 10 can be accommodated by Pascal's Triangle and a hand calculator. More extensive computational resources need to be found for n greater than 10. Tables of the Binomial cumulative distribution exist to handle such situations.

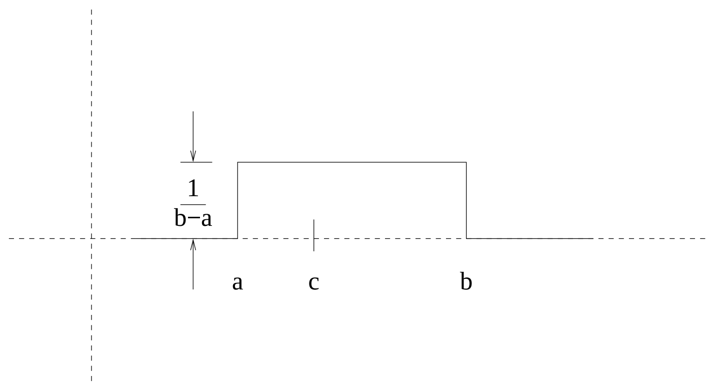
Problem 3. *Find* $P(T = 6)$ *for* $T \sim b(15, .3)$.

We need to refer to the Binomial table for n=15. From the table, it is seen that $P(T \leqslant 6) = .869$ *and* $P(T \leqslant 5) = .722$*, so that* $P(T = 6) = P(T \leqslant 6) - P(T \leqslant 5) = .147.$

9 Continuous Random Variables

9.1 Uniform $U(a, b)$

The best way to describe the Uniform distribution for the continuous random variable is with a picture of its probability density function as shown here.



The area under the curve of any probability density function is always one. The Uniform distribution is no exception. The area under it must be equal to one; therefore, the area of the rectangle seen in the probability density function must be equal to one, that is: $(b-a)\frac{1}{b-a} = 1$.

As in the discrete case where the sum of the probabilities is $\sum_k P(X = k) = 1$, in the continuous case, the area under the curve is 1. Because continuous random variables involve intervals and not discrete ranges, the study of continuous random variables falls into the realm of infinities or continua. As a consequence, the area under the probability density function replaces the summation of probabilities. That change in focus from calculating the sum of probabilities to the calculation of areas under a curve depends on the principles of integral calculus. On account of continuous random variables, the mathematics of statistics becomes very sophisticated very quickly. However, since the Uniform distribution is so simple, the analysis of it can be done without resorting to calculus.

Definition 20. *If $X \sim U(a, b)$, then*

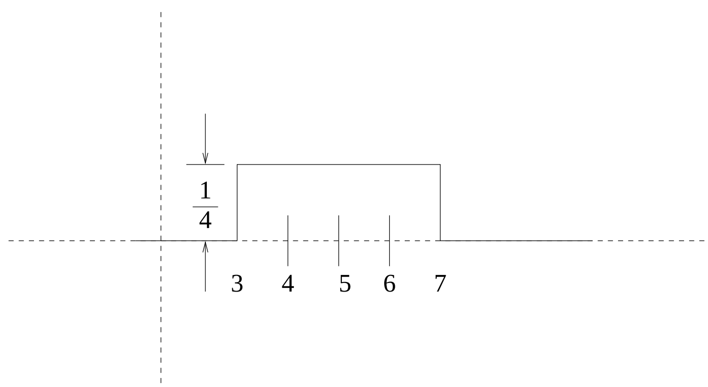
$$P(X \leqslant c) = \begin{cases} 0 & if \quad c \leqslant a \\ \frac{c-a}{b-a} & if \quad a \leqslant c \leqslant b \\ 1 & if \quad c \geqslant b \end{cases}$$

It is not necessary to memorize formulas like this one because it is better to rely on pictures which depict the essence of the situation. It is obvious, for instance, by looking at the picture of the probability density function for a Uniform distribution that the area under a point is zero. This fact marks one of the significant differences between discrete and continuous random variables. If a random variable, Y, is continuous, then $P(Y = c) = 0$ for any c. Whereas, if X is a discrete random variable like the discrete Uniform random variable, then $P(X = k) = 1/n \neq 0$.

If $X \sim U(a, b)$, then it has an expected value and a variance. The expected value and the variance of X are obtained by means of integral calculus. The respective formulas are: $E[X] = \frac{a+b}{2}$ and $var(X) = \frac{(b-a)^2}{12}$. In the special case when $X \sim U(0, 1)$, $E[X] = \frac{1}{2}$ and $var(X) = \frac{1}{12}$.

Uniform U(a,b)
$E[X] = \frac{a+b}{2}$ $var(X) = \frac{(b-a)^2}{12}$

Example 21. *Suppose $W \sim U(3, 7)$. The first step in addressing a random variable is to draw a picture of its distribution. The height of the rectangle must be the reciprocal of its length because the area under the curve must be one.*



Having drawn a picture, the following probabilities are easy to deduce.

1. $P(W \leqslant 2) = 0$
2. $P(W \leqslant 5) = 2(\frac{1}{4}) = \frac{1}{2}$
3. $P(W > 6) = \frac{1}{4}$
4. $P(4 \leqslant W \leqslant 6) = \frac{1}{2}$
5. $E[W] = \frac{3+7}{2} = 5$

6. *$var(W) = \frac{(7-3)^2}{12} = \frac{16}{12} = \frac{4}{3}$*

7. *Find c such that $P(W \leqslant c) = .6$ The answer is: c=5.4*

9.2 Mathematical Interlude

Associated with a continuous random variable is a probability density function. In the discrete case, it is called the probability mass function. For a continuous random variable, it is called the probability density function.

Definition 21. *Denote the probability density function, pdf, as f(x).*

Henri Lebesgue
1875-1941

Felix Edouard Justin Emile Borel
1871-1956

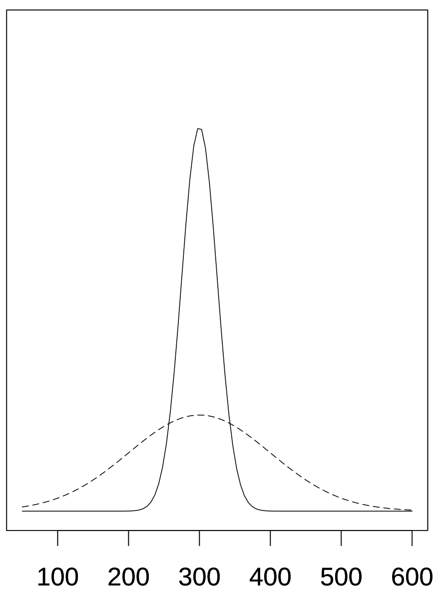
Discrete and continuous random variables are different. It is easy to understand the development of a discrete random variable, but that is not the case with continuous random variables. It was not until the 1930's when theoretical advances made by Henri Lebesgue augmented the work of Emile Borel that mathematicians finally got a firm grasp on the interpretation of continuous random variables. Then both discrete and random variables came under the same mathematical discipline known as measure theory. A probability as we already know is the relative measure of an event to the sample space. Because the continuous random variables involve infinitisimals and expressions extending to infinity, the use of integral calculus is unavoidable. However, close parallels between discrete and continuous random variable do exist like those given in Table 2, so that the method for computing probabilities of continuous random variables follows along the same line of reasoning as in the procedure for the discrete case except we will extensively use tables

Table 2

	For Discrete Random Variable	For Continuous Random Variable
CDF	$P(X \leqslant c) = \sum_{\substack{over\ all \\ values\ of\ X \leqslant c}} P(X = k)$	$P(X \leqslant c) = \int_{-\infty}^{c} f(t)dt$
Expected Value	$E[X] = \sum_{\substack{over\ all \\ values\ of\ X}} kP(X = k)$	$E[X] = \int_{-\infty}^{\infty} tf(t)dt$
Variance	$var(X) = \sum_{\substack{over\ all \\ values\ of\ X}} (k - E[X])^2 P(X = k)$	$var(X) = \int_{-\infty}^{\infty} (t - E[X])^2 f(t)dt$

in lieu of calculus. Underlying these tables is a massive foundation of technically difficult numerical methods which rely not surprisingly on a solid understanding of advanced calculus.

9.3 Normal Distribution $N(\mu, \sigma^2)$



The Uniform distribution is the simplest distribution among those for a continuous random variable, but it is not the easiest one to work with. Ironically, the Normal distribution defined by an intimidating formula of its probability density function is the nicest of all distributions. The Normal distribution was coined by Jules Henri Poincare, but it is often referred to as the Gaussian distribution by engineers and scientists in honor of Friedrich Gauss who invented the Normal distribution for his newly developed method of least squares, a numerical technique which will be discussed in detail later during our study of linear models.

The equation of the probability density function (pdf) of the Normal distribution, $N(\mu, \sigma^2)$, is: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. The graph of two Normal pdf's are shown in the figure above. Both of them have the same value of μ which is 300. This parameter determines the location of the distribution. The Normal distribution is symmetrical about the mean; therefore, μ is the center of mass of the Normal distribution. The other parameter, σ^2, determines the shape of the distribution. The larger σ^2 is, the flatter the shape of the Normal distribution. The dashed curve represents a $N(300, 10000)$ while the curve with the solid line represents a $N(300, 625)$.

The symmetry, graceful curvature, and unique shape which our eyes immediately perceive reveal some of the seemingly boundless secrets of the Normal distribution. Francis Galton, a early pioneer of modern statistics, wrote a eloquent description of the Normal Distribution in chapter V of his book, *Natural Inheritence*, published in 1889:

> *Order in Apparent Chaos.* - I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the " Law of Frequency of Error." The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along. The tops of the marshalled row form a flowing curve of invariable proportions ; and each element, as it is sorted into place, finds, as it were, a pre-ordained niche, accurately adapted to fit it. If the measurement at any two specified Grades in the row are known, those that will be found at every other Grade, except towards the extreme, ends, can be predicted in the way already explained, and with much precision,

While our intellect is immediately drawn to the equation of the Normal distribution, we notice that its exponent is raised to an exponent. As curious as that might be, we cannot help but notice the presence of two peculiar numbers in an enigmatic relationship.

The history of π makes it the most famous of all the mathematical constants. A fascinating account of π appears in the book, *A History of* π, written by Petr Beckmann, a man who fled Czechoslovakia in 1968 and settled in the United States. An electrical engineer by profession, Beckmann's story of π nicely describes the relation of π with the history of mathematics. By definition, π is the ratio of the circumference of a circle to its diameter, that is: $\pi = \frac{C}{D} = 3.141592653\ldots$. A precise decimal approximation of π essentially eluded mathematicians until the advent of numerical methods which were made possible by the invention of calculus.

Every year seems to bring news of the discovery of many more digits in the decimal expansion of π which reached 12,100,000,000,050 digits on 28 December 2013 [1] . In spite of the amazing pursuit of finding ever more digits, no discernible pattern in the digits has been found. Every civilization knew of π, but it was Greek geometry which addressed π theoretically by none other than Archimedes who first proved that the area of a circle is: πr^2. His crowning achievement and the one which he insisted be inscribed on his grave marker symbolically by a sphere and right cylinder was the discovery that the volume of a sphere is: $\frac{4}{3}\pi r^3$. Who knows, had he not been slain by a Roman soldier, Archimedes might have discovered calculus 1,800 years before Newton. The universal fame of this fundamental constant of mathematics which we call π never diminishes.

Johann Carl Friedrich Gauss
1777-1855

Leonhard Euler
1707-1783

Another fundamental constant is $e = 2.718281828\ldots$. Its origins lie in differential calculus where it was discovered and named by Leonhard Euler. One by one, as functions succumbed to the ambitions of mathematicians to differentiate them in the early days of calculus, the trigonometric functions and the logarithmic function were the simplest non-algebraic functions to receive attention and were conquered by the great mathematicians. In the process of differentiating the logarithmic function, $\frac{d\,log_a(x)}{dx} = \frac{log_a(e)}{x}$, Euler found the constant e to be $\lim_{n\to\infty}(1+\frac{1}{n})^n$. It is much easier to find the decimal expansion of e than that of π. In many ways, e is a nice number, unlike the stubborn π, though it is π which for some inscrutable reason enjoys general fascination. Yet, like π, e appears everywhere in mathematics.

What is truly remarkable is that the number π which originates in geometry and the

[1]See: http://en.wikipedia.org/wiki/Chronology_of_computation_of_π

number e which originates in calculus should come sublimely together in probability in the equation of the Normal distribution without any discernible reason. Why of all places should such a meeting of two basic constants even occur is a question which drives mathematicians to marvel at the beauty of mathematics. It incites them to ponder the perennial question: What constitutes a mathematical discovery? Do things like the Normal distribution exist ever since the creation of the universe waiting to be discovered by someone or are mathematical discoveries only a figment of man's imagination? The Normal distribution is indeed fundamental. It is the bread and butter of the statistician. Almost nothing of general practical importance in statistics does not depend on or cannot be expressed in terms of or approximated to any degree by the Normal distribution. Its meaning runs deep; its importance is unsurpassed.

The notation which is used to identify a Normal distribution is: $N(\mu, \sigma^2)$. That the two parameters, μ and σ^2, are used to characterize the Normal distribution and, also, to denote the population mean and population variance is not merely co-incidental. We will soon see that a clever use of the Normal distribution will make it possible to describe almost any set of experimental data.

If $X \sim N(\mu, \sigma^2)$, then X has an expected value and a variance. By means of the definition of the expected value for a continuous random variable as shown in Table 2 and the non-trivial application of the rules of advanced calculus, $E[X] = \mu$ and for the variance, $var(X) = \sigma^2$. That is, E[X] and var(X) in the world of probability appear to correspond to the population mean and population variance in the world of statistics. Indeed, one may judge the resemblance of a histogram and a probability distribution not only by looking at them, but one can test analytically whether the probability distribution adequately agrees with a histogram. If the center of mass of a random variable, $E[X]$, agrees with the center of mass, μ, of a population according to some specified criterion, then a bridge will have been made between the theory of probability and the data of statistics. In that case, the association of a probability distribution with a histogram of the data will be deemed defensible, and the job of the statistician who is hired to analyze the problem will essentially come to an end.

Normal N(μ,σ^2)
$E[X] = \mu$ $var(X) = \sigma^2$

Two important properties of the Normal distribution are:

1. If $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$ for the two arbitrary constants, a and b. In particular, let $a = \frac{1}{\sigma}$ and $b = -\frac{\mu}{\sigma}$, then $aX + b = \frac{X-\mu}{\sigma}$; moreover, $a\mu + b = 0$, and $a^2\sigma^2 = 1$, hence $\frac{X-\mu}{\sigma} \sim N(0, 1)$.

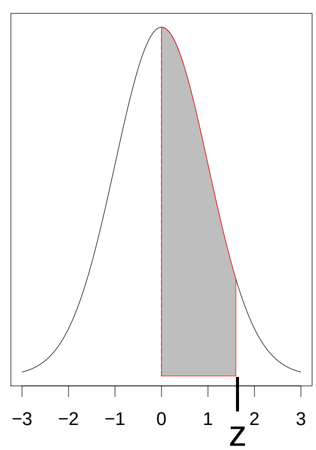
2. A set of independent Normal random variables can be used to approximate other random variables.

Definition 22. $\frac{X-\mu}{\sigma}$ *is called the* ***population z-score*** *of* X.

Definition 23. $\frac{x_i-\bar{x}}{s}$ *is called the* ***sample z-score*** *of* x_i.

Both z-scores are examples of scaling. There are an infinite number of possible Normal distributions, but all Normal random variables can be transformed into the Standard Normal distribution, $N(0,1)$ by means of the z-score, and it suffices, therefore, to have only one table for the Normal distribution like the one given in Appendix 12.

10 Computations Using Continuous Random Variables



Let $X \sim N(0,1)$. A picture of this distribution will prove to be an invaluable aid in solving problems.

Observations 2.

- *The area under the curve is 1.*
- *The area from* $-\infty$ *to 0 is 1/2 which implies that* $P(X < 0) = .5$.
- *Obviously,* $P(X \leqslant z) + P(X > z) = 1$*; therefore,* $P(X > z) = 1 - P(X \leqslant z)$.
- *By symmetry,* $P(X \leqslant -z) = P(X \geqslant z)$.
- *Recall that* $P(X = z) = 0$ *for any* z*, because* X *is a continuous random variable.*

Other than trivial problems, numerical methods are necessary for producing decimal numbers of a probability. As it is always possible to transform $X \sim N(\mu, \sigma^2)$ into a $N(0,1)$ by means of a z-score, only one table is sufficient to find probabilities for the Normal distribution. The table is constructed to provide the area under the probability density function between 0 and z; therefore, the area to the left of 0 needs to be considered by adding .5 when it is appropriate. These ideas will be made clear by the following example.

Example 22. *Given that $X \sim N(0,1)$:*

1. $P(X < .1) = .5 + .0398 = .5398$

2. $P(X \leqslant 1) = .5 + .3413 = .8413$

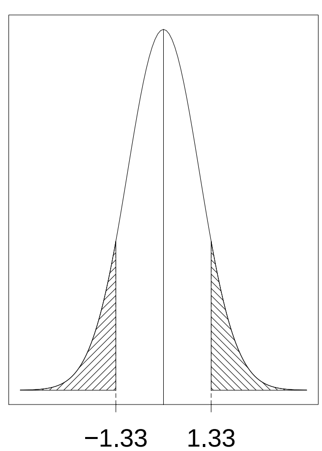
3. $P(X \geqslant 1.5) = 1 - P(X < 1.5) = 1 - (.5 + .4332) = .0668$

4. $P(X \leqslant -.5) = P(X \geqslant .5) = 1 - P(X \leqslant .5) = 1 - (.5 + .1915) = .3085$

Example 23. *Suppose $Y \sim N(6, 2.25)$. Immediately, we know that the mean is $\mu = 6$ and the variance is $\sigma^2 = 2.25$, i.e. $\sigma = 1.5$; therefore, the necessary ingredients to find the z-score are readily available. What is $P(Y \leqslant 4)$?*

$$\begin{aligned} P(Y \leqslant 4) &= P(Y - 6 \leqslant 4 - 6) \\ &= P\left(\frac{Y-6}{1.5} \leqslant \frac{4-6}{1.5}\right) \\ &= P(z \leqslant -1.333) \end{aligned}$$

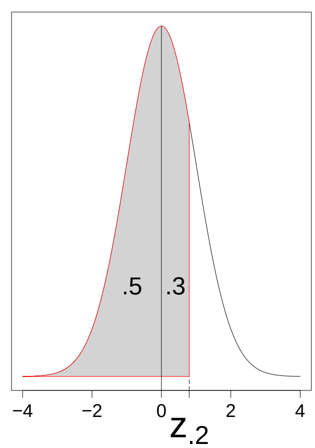
But the table of probabilities provided in Appendix 12 does not provide probabilities for negative values of z. However, the area under the curve up to -1.33 is the same area to the right of 1.33 in accordance with our earlier observations of the symmetry of the Normal distribution which implies that $P(X \leqslant -z) = P(X \geqslant z)$.



$$
\begin{aligned}
P(Y \leqslant 4) &= P(z \leqslant -1.333) \\
&= P(Z \geqslant 1.333) = 1 - P(Z \leqslant 1.3333) = 1 - (.5 + .4082) \\
&= .0918
\end{aligned}
$$

The converse of finding a probability is to find a z which produces a specified probability.

Example 24. *Find x_0 such that $P(X \leqslant x_0) = .8$ when $X \sim N(0,1)$. To help correlate x_0 and .8 with a z-score, we will use the notation, $z_{.2}$. Both numbers, x_0 and $z_{.2}$, are the same. $z_{.2}$ gives the connotation of a z-score such that the area to the right of it is .2 or equivalently the area to the left of it is .8 as depicted in the following picture.*



The picture says that $P(X \leqslant z_{.2}) = .8 = .5 + .3$. That number, $z_{.2}$, which produces an area under the curve of .3 between 0 and itself corresponds to that z in the table of probabilities which gives a value of .30000 in the body of the table. Although there is no such number in the body of the table as .3000 the number that comes closest to it is used instead. That number in the body of the table is .29955, and it corresponds to a z of .84. In conclusion, $x_0 = z_{.2} = .84$. It is always prudent to check one's answer. To that end, $P(X \leqslant .84) = .29955$; the answer is, therefore, correct.

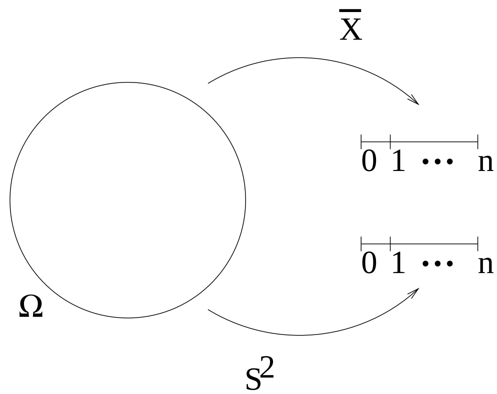
11 Sampling Distributions

Question 1. *Which of the following, if any, are random variables?*

$$
\mu = \frac{\sum\limits_{i \in \mathcal{P}} x_i}{N}, \qquad \bar{x} = \frac{\sum\limits_{i \in \mathcal{S}} x_i}{n}, \qquad \sigma^2 = \frac{\sum\limits_{i \in \mathcal{P}} (x_i - \mu)^2}{N}, \qquad \text{or } s^2 = \frac{\sum\limits_{i \in \mathcal{S}} (x_i - \bar{x})^2}{n-1}.
$$

A random variable maps an outcome of a sample space to a number. In the case of μ and σ^2, the sample space consists of only one outcome namely, $\mathcal{P}$. In a sense, μ and

σ^2 are degenerate random variables; they are constants. On the other hand, $\bar{x}$ and s^2 are different for each draw of a sample. The sample space for them consists of every possible sampling of elements of $\mathcal{P}$. For each sample, $\bar{x}$ maps it to a number and s^2 maps the sample to another number. Both $\bar{x}$ and s^2 are random variables. The schematic diagram shown below illustrates the mapping of $\bar{x}$ and s^2 from the same outcomes to different numbers.

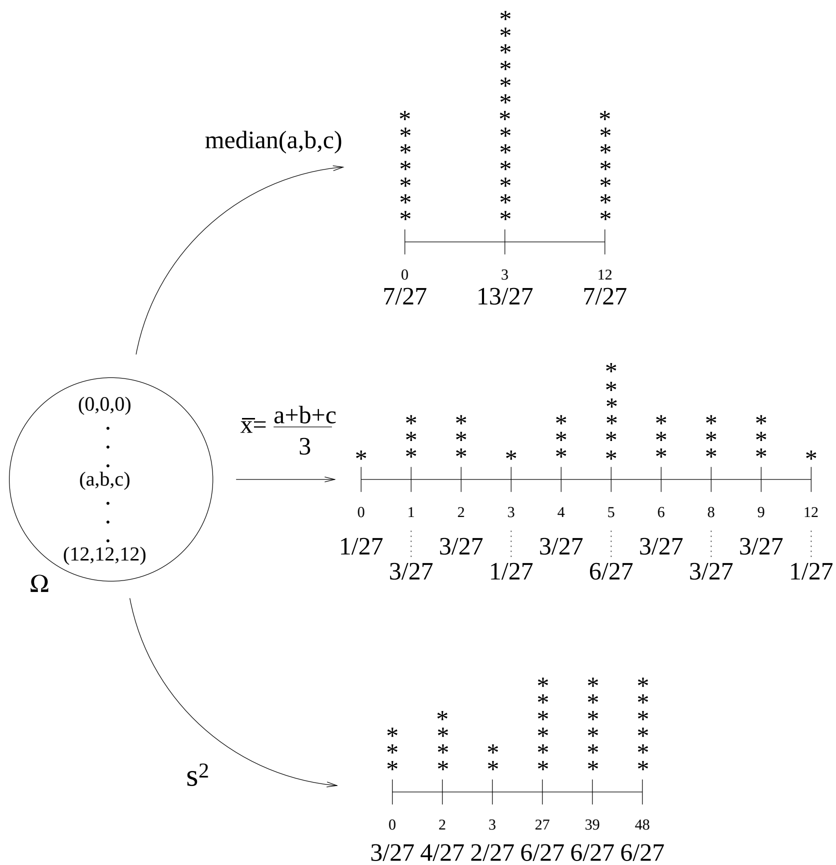


Not only are $\bar{x}$ and s^2 random variables, but any mapping of a sample to a number is a random variable like the sample median or the sample 1^{st} quartile or the sample range. Associated with a random variable is a probability distribution. There is one for $\bar{x}$ and a different one for s^2. The probability distribution which is associated with a sampling random variable is something called a sampling distribution simply, in order to emphasize its association with a sample.

To illustrate the concept of a sampling distribution, consider the sample space of outcomes in which an outcome consists of a triple. Any place in the triple can be filled with either a 0, 3, or a 12. As such, the sample space is:

$$\Omega = \left\{ \begin{array}{lll} (0,0,0) & (3,0,0) & (12,0,0) \\ (0,0,3) & (3,0,3) & (12,0,3) \\ (0,0,12) & (3,0,12) & (12,0,12) \\ (0,3,0) & (3,3,0) & (12,3,0) \\ (0,3,3) & (3,3,3) & (12,3,3) \\ (0,3,12) & (3,3,12) & (12,3,12) \\ (0,12,0) & (3,12,0) & (12,12,0) \\ (0,12,3) & (3,12,3) & (12,12,3) \\ (0,12,12) & (3,12,12) & (12,12,12) \end{array} \right\}$$

Define $\bar{x} = \frac{a+b+c}{3}$ where a, b, and c are the places in any triple (a,b,c). $\bar{x}$ maps an outcome to the average of its three members. Define $s^2 = \frac{(a-\bar{x})^2+(b-\bar{x})^2+(c-\bar{x})^2}{2}$, and the median as the median of a, b, and c. Each random variable has a set of possible values. For $\bar{x}$, the possible values are: $\{0, 1, 2, 3, 4, 5, 6, 8, 9, 12\}$; for s^2, the possible values are: $\{0, 2, 3, 27, 39, 48\}$; for the median, the possible values are: $\{0, 3, 12\}$. Associated with each of these three random variables is a probability distribution; they are shown in the accompanying diagram. None of the distributions is a common distribution which we know by a name, nevertheless, the diagram tells us everything we need to know about the distributions of $\bar{x}$, s^2, and the median. From the diagram, for instance, it is can be seen that $P(\bar{x} = 4) = \frac{3}{27}$. Similarly, $P(median = 3) = \frac{13}{27}$ and $P(s^2 = 27) = \frac{6}{27}$. It is clear that the sample mean, sample variance, and the sample median are random variables and that they each have a different probability distribution.



Consider the random variable, $\bar{x}$. It has an expected value and a variance, that is:

$$E[\bar{x}] = 0(\frac{1}{27}) + 1(\frac{3}{27}) + \cdots + 9(\frac{3}{27}) + 12(\frac{1}{27}) = 5$$

and

$$var(\bar{x}) = (0-5)^2\frac{1}{27} + \cdots + (12-5)^2(\frac{1}{27}) = \frac{78}{9}$$

On a different tack, it is enlightening to examine these probability distributions from another perspective in which each is explained by three independent constituents corresponding to each place of a triple. Each place in a triple consists of either a 0, 3, or a 12, and they occur equally likely. Each place of a triple can be associated with its own sample space: $\Omega_1 = \{0, 3, 12\}$, $\Omega_2 = \{0, 3, 12\}$, and $\Omega_3 = \{0, 3, 12\}$. Because 0, 3, and 12 are put into each place of a triple without prejudice, we may say that the event that a 0, 3, or 12 appears in place 1 and the event that 0, 3, or 12 appears in place 2 and the event that 0, 3, or 12 appears in place 3 are equally likely. Let X_1 be a random variable such that it maps Ω_1 into 0, 3, and 12. By assumption of equally likely outcomes, $P(X_1 = 0) = P(X_1 = 3) = P(X_1 = 12) = \frac{1}{3}$. The random variable, X_1, has an expected value, namely $E[X_1] = 0(\frac{1}{3}) + 3(\frac{1}{3}) + 12(\frac{1}{3}) = 5$. What is characteristic of X_1 is also characteristic of X_2 and X_3. All three random variables have the same properties. How each slot in the triple is filled is independent of how the others are filled; therefore, X_1, X_2, and X_3 are independent random variables. They are identically distributed which implies that $E[X_1] = E[X_2] = E[X_3] = 5 = \mu$ and $var(X_1) = var(X_2) = var(X_3) = (0-5)^2(\frac{1}{3}) + (3-5)^2(\frac{1}{3}) + (12-5)^2(\frac{1}{3}) = \frac{78}{3} = \sigma^2$.

Definition 24. *i.i.d. means **independent identically distribution**.*

Earlier, we found that $E[\bar{x}] = 5$ and $var(\bar{x}) = \frac{78}{9}$ by direct computation. By definition, $\bar{x} = \frac{x_1+x_2+x_3}{3}$, and X_1, X_2, and X_3 are i.i.d. with expected value, μ, and variance, σ^2, so that in light of the preceding paragraph, $E[\frac{x_1+x_2+x_3}{3}] = E[\bar{x}] = 5 = \frac{5+5+5}{3} = \frac{E[x_1]+E[x_2]+E[x_3]}{3} = \frac{3\mu}{3} = \mu$. In other words, $E[\frac{x_1+x_2+x_3}{3}] = \frac{E[x_1]+E[x_2]+E[x_3]}{3} = \frac{3E[x_i]}{3} = E[x_i]$. Likewise,

$$\begin{aligned} var\left(\frac{x_1+x_2+x_3}{3}\right) = var(\bar{x}) &= \frac{78}{9} = \frac{\frac{78}{3}+\frac{78}{3}+\frac{78}{3}}{3^2} \\ &= \frac{var(X_1)+var(X_1)+var(X_1)}{3^2} = \frac{3\sigma^2}{3^2} = \frac{\sigma^2}{3} \end{aligned}$$

The expressions of $E[\bar{x}]$ and $var(\bar{x})$ which are written in terms of X_1, X_2, and X_3 suggest the formulation of a general theorem.

lemma 1. *If $X_1, X_2, \ldots, X_n$ are i.i.d., then $E[X_1] = E[X_2] = \ldots = E[X_n]$ and $var(X_1) = var(X_2) = \ldots = var(X_n)$.*

Theorem 8. *If X and Y are random variables, then E[X+Y]=E[X]+E[Y].*

Theorem 9. *If a and b are constants, then E[aX+b]=aE[X]+b and var(aX+b)=a^2var(X).*

Theorem 10. *If X and Y are independent random variables, then var(X+Y)=var(X)+var(Y).*

Theorem 11. *If $X_1, X_2, \ldots, X_n$ are i.i.d. each with mean μ and variance σ^2, and $\bar{x} = \frac{X_1+\cdots+X_n}{n}$, then*

$$E[\bar{x}] = \mu \text{ and } var(\bar{x}) = \frac{\sigma^2}{n}$$

Proof. $E[\bar{x}] = E[\frac{X_1+X_2+...+X_n}{n}] = \frac{E[X_1+...+X_n]}{n} = \frac{E[X_1]+...+E[X_n]}{n} = \frac{\mu+...+\mu}{n} = \frac{n\mu}{n} = \mu$
$var(\frac{X_1+X_2+...+X_n}{n}) = \frac{var(X_1)+var(X_2)+...+var(X_n)}{n^2} = \frac{\sigma^2+...+\sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$ ∎

We had observed that X_1, X_2, and X_3 are identically distributed. If it is assumed that they are independent random variables then they are i.i.d. random variables each with μ=5 and $\sigma^2 = \frac{78}{3}$. Accordingly, by Theorem 11, $E[\bar{X}] = 5$ and $var(\bar{X}) = \frac{\frac{78}{3}}{3} = \frac{78}{9}$ which are in exact agreement with the values which were obtained directly from definition. Getting the same answers by means of two different ways illustrates a rule of problem solving. There are usually two ways to solve a problem: a short and easy way and a long and difficult way. Computing $E[\bar{X}]$ and $var(\bar{X})$ directly from the definitions of expected value and variance is the long and difficult way. Resorting to a theorem like Theorem 11 is the short and easy way. Even though both ways produce the same answers, who, when time is precious and patience is short as it often happens during an examination, would not choose to use the theorems?

$\bar{x}$	median	s^2
$E[\bar{x}] = 5$	$E[median] = \frac{123}{27}$	$E[s^2] = \frac{698}{27}$
$median(\bar{x}) = 5$	$median(median) = 3$	$median(s^2) = 27$
$var(\bar{x}) = \frac{78}{9}$	$var(median) = \frac{112}{3}$	$var(s^2) = \frac{251462}{729}$

In conclusion, a summary of the results gotten from directly applying the basic definitions of expected value and variance for the three sampling random variables, $\bar{x}$, *median*, and s^2, is given above.

Example 25. *Let $X_i \sim b(100, \frac{1}{5})$, denote $\bar{x} = \frac{X_1+X_2+X_3+X_4}{4}$ and suppose x_i's are independent; i.e. they are i.i.d.*

1. $E[X_1] = E[X_2] = E[X_3] = E[X_4] = np = 100\frac{1}{5} = 20 = \mu$
2. *By Theorem 11,* $E[\bar{x}] = \mu = 20$ *or*
   $E[\frac{X_1+X_2+X_3+X_4}{4}] = \frac{E[X_1]+E[X_2]+E[X_3]+E[X_4]}{4} = \frac{20+20+20+20}{4} = 20$
3. $var(X_i) = npq = 100\frac{1}{5}\frac{4}{5} = 16 = \sigma^2 \quad i = 1, 2, 3, 4$
4. *By Theorem 11,* $var(\bar{x}) = \frac{\sigma^2}{n} = \frac{16}{4} = 4,$ *or*
   $var(\frac{X_1+X_2+X_3+X_4}{4}) = \frac{var(X_1)+var(X_2)+var(X_3)+var(X_4)}{16} = \frac{4[100(\frac{1}{5})(\frac{4}{5})]}{16} = 4.$

12 Estimation of Parameters

Suppose three fair coins are flipped and the number of heads that appear are counted by the random variable, X, then $X \sim b(3, .5)$. The coins might not be fair but actually

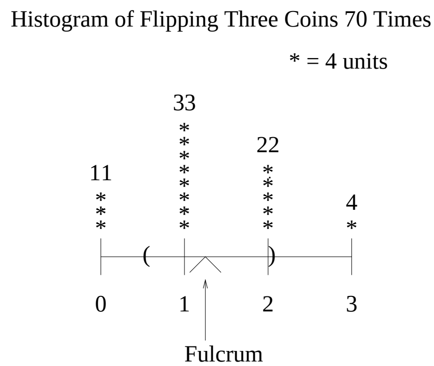
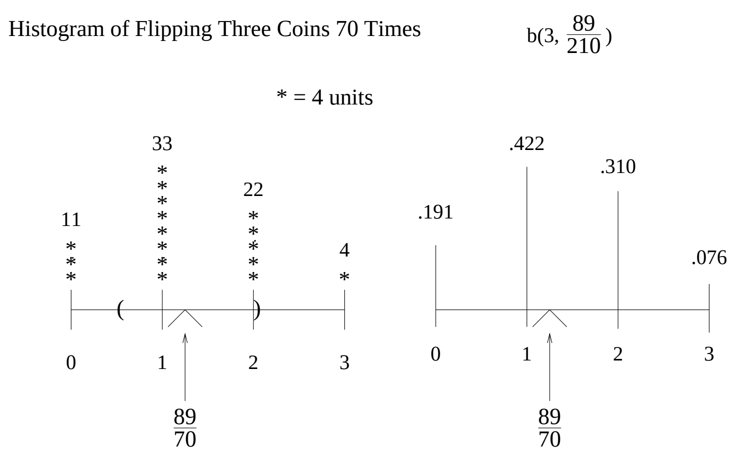


Figure 4

loaded ones. The probability of getting a head, in this alternate experiment, is unknown but denote it by p, so that $X \sim b(3, p)$. Based on the data obtained from the experiment of flipping three coins 70 times of unknown probability of getting a head, the histogram shown on the left was constructed. This is the same histogram which appears in Figure 12.



The histogram looks like a Binomial distribution. If that is the case, then what Binomial distribution comes closest to fitting the histogram? To answer that question, we hypothesize that the random variable of counting the number of heads be distributed as a generic Binomial distribution like $X \sim b(\nu, p)$ with yet to be determined parameters, ν and p. If the probability distribution is supposed to describe the data which we see presented in the form of a histogram, then it should have, at least, the same center of mass as the one for the histogram. That Binomial distribution which comes closest in matching the histogram will be that one for which $E[X] = \bar{x}$. For $X \sim b(\nu, p)$, $E[X] = \nu p$; therefore, $\nu p = \bar{x}$ is the necessary condition that we are imposing to find the best Binomial distribution for the data. Simply solving for p gives: $\hat{p} = \frac{\bar{x}}{\nu}$. This particular value of p is an empirically derived number from real data; it is not a probability which is something theoretical. In order to avoid confusion over this matter, statisticians put a *hat* over p to signify that this p is an estimate based on data of the theoretical p.

We know already that $\nu = 3$ because three coins were flipped and $\bar{x} = \frac{89}{70}$; therefore, $\hat{p} = \frac{\frac{89}{70}}{3} = \frac{89}{210} \approx .4238$. A picture of $b(3, \frac{89}{210})$ is shown to the right of the histogram. It looks as if it makes a good fit with the histogram. But a closer inspection reveals that $b(3, \frac{89}{210})$ does not match the histogram exactly. It does, out of all possible Binomial distributions, fit the histogram the best, but it is not an exact match as the entries in the following table prove.

k	Observed Frequencies	Estimated Probabilities from $b(3, \frac{89}{210})$
P(X=0)	$\frac{11}{70} = .157$	$\binom{3}{0}\hat{p}^0(1-\hat{p})^3 = \frac{1771561}{9261000} = .191$
P(X=1)	$\frac{33}{70} = .471$	$\binom{3}{1}\hat{p}^1(1-\hat{p})^2 = \frac{3909147}{9261000} = .422$
P(X=2)	$\frac{22}{70} = .314$	$\binom{3}{2}\hat{p}^2(1-\hat{p})^1 = \frac{2875323}{9261000} = .310$
P(X=3)	$\frac{4}{70} = .057$	$\binom{3}{3}\hat{p}^3(1-\hat{p})^0 = \frac{704969}{9261000} = .076$

There might be many reasons for the discrepancy. The most obvious one is that some other probability distribution is a better candidate to describe the data. It might be such a novel distribution that it might not have been discovered yet. Perhaps the discrepancy is due to insufficient number of flips of the coin. Our intuition tells us that the more experimental data, the better the estimates, so that with enough flips the histogram and the Binomial distribution will converge to the same thing. Perhaps the flips were not performed independently of each other. Perhaps the person who did the flipping is not a good flipper of coins, and so on.

Ultimately, we would like to associate a probability distribution with each of the descriptive statistics, in order to explain the data. An even better result would be to find a random variable whose expected value equals the population mean and whose variance equals the population variance. Such a quest is almost impossible to do if one is restricted to a one parameter distribution like the Binomial distribution. If an exact fit is not possible in general, then the next best thing to do is to find a probability distribution that comes close enough to the histogram, so that it does adequately describe the data to suit our needs. What is close or not close is an issue that will be discussed in the topic of testing hypotheses, but, until then, we will recognize the remarkable discovery that there is one distribution which provides the means of estimating both the mean and the variance and which can adequately imitate a given histogram most of the time. That distribution is none other than the renowned Normal distribution.

Suppose $X_i \sim N(\mu, \sigma^2)$ and that it is hypothesized that X_i represents the i^{th} measurement, x_i, which is taken from an actual experiment. If that assertion is true, then because $E[\bar{X}] = \mu$ we may assert that $\hat{\mu} = \bar{x}$ where $\bar{x}$ is the sample mean. In other words, the assertion implies that the sample mean is an estimate of the population mean. Similarly, the sample variance, s^2, is an estimate of the population variance σ^2. The key ingredient for making these assertions is that the Normal distribution via $X_i \sim N(\mu, \sigma^2)$ provides a bridge, though mathematically complicated to justify, between the world of probability signified by (μ, σ^2) and the real world signified by $(\bar{x}, s^2)$. The Normal distribution is truly a remarkable distribution. No other distribution is so nice.

Not all estimators do a good job in estimating a parameter of a probability distribution. An estimator might be biased. For example, a claim might be made that $\hat{\mu} = 3\bar{x} + 5$. The question arises whether that estimator of μ is a good one in the long run. It might be good perhaps once or twice but never again. Is it good in the long run? That is, in expectation, will it equal the population mean? Taking the expected value produces: $E[\hat{\mu}] = E[3\bar{x} + 5] = 3E[\bar{x}] + 5$ by Theorem 9. By Theorem 11, $3E[\bar{x}] + 5 = 3\mu + 5 \neq \mu$. That estimator is definitely biased. This leads to an important definition:

Definition 25. *If $\hat{\Theta}$ is an estimator of Θ, then $\hat{\Theta}$ is called an* ***unbiased estimator*** *of Θ, if $E[\hat{\Theta}] = \Theta$.*

If X_i are i.i.d. $N(\mu, \sigma^2)$ and $\bar{x} = \frac{X_1+X_2+\cdots+X_n}{n}$, then $E[\bar{x}] = \mu$; hence $\bar{x}$ is called an unbiased estimator of μ. Also, $E[s^2] = \sigma^2$; hence s^2 is called an unbiased estimator of σ^2. It is only now after a good deal of discussion on the subject of probability that a justification can be given for defining s^2 the way we did back in Chapter 1. Depending on the author, s^2 might be defined to be: $s^2 = \frac{\sum_{i\in S}(x_i-\bar{x})^2}{n}$ then $E[s^2] = \frac{n-1}{n}\sigma^2$ which implies that this definition of s^2 makes it a biased estimator of σ^2. That is why n-1 is put in the denominator of our definition of s^2, in order to make it an unbiased estimator of σ^2.

If X_i are i.i.d. $N(\mu, \sigma^2)$, then it can be proved that the z-score of $\bar{x}$ is distributed as a Standard Normal distribution; i.e. $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,\ 1)$.

Example 26. *Let $X_i \sim N(5, 9)$ for i=1, 2, 3, 4, and 5 be independent random variables.*

1. $E[X_i] = 5$.
2. $var(X_i) = 9$.
3. $E[\bar{x}] = 5$.
4. $var(\bar{x}) = \frac{\sigma^2}{5} = \frac{9}{5}$.

Suppose, on the other hand, that X_i are independent identically distributed but with an unknown distribution. Such would be the case in any actual experiment like those involving college examination scores, weights of tomatoes, number of home runs, speed of asteroids, or whatever. How any of these random variables might be distributed is completely unknown. Nevertheless, the Central Limit Theorem says that it is sufficient that the random variables be i.i.d. for the sample z-score to converge to $N(0,1)$ as the number of observations goes to infinity.

Theorem 12. *(Central Limit Theorem) If $X_1,\ X_2,\ \ldots\ ,X_n$ are i.i.d. and $\bar{x} = \frac{X_1+\cdots+X_n}{n}$ then*

$$\frac{\bar{x} - E[\bar{x}]}{\sqrt{var(\bar{x})}} = \textit{sample z-score} \quad \rightarrow \quad N(0,\ 1) \quad \textit{as } n \rightarrow \infty$$

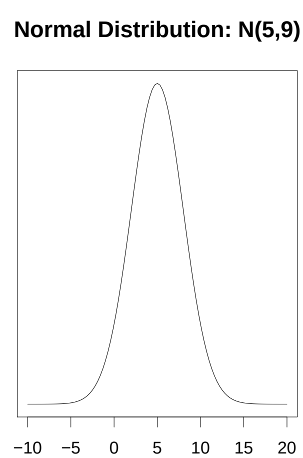
This theorem is another indication of the amazing versatility of the Normal distribution. Regardless of how X_i is distributed, so long as they are i.i.d., the sample z-score will converge to $N(0,1)$ as the number of observations increase to infinity.

A celebrated story relating the utility of the Central Limit Theorem to the behavior of the masses was given be Francis Galton.

> In 1906, Galton visited a livestock fair and stumbled upon an intriguing contest. An ox was on display, and the villagers were invited to guess the animal's weight after it was slaughtered and dressed. Nearly 800 gave it a go and, not surprisingly, not one hit the exact mark: 1,198 pounds. Astonishingly, however, the average of those 800 guesses came close very close indeed. It was 1,197 pounds.

It is this capacity of the Central Limit Theorem to explain social behavior of a class of people which makes sociology possible.

Example 27. *We are given a set of data which was obtained from a sample of size 10. Suppose someone with prior knowledge about the situation informed us that each observation represented the realized value of a random variable which is distributed as a $N(5,9)$.*



1. *The statement of the problem implies that the Normal distribution cited in the problem adequately explains the population. Consequently, without seeing the data nor doing any calculations, the population mean is:* $\mu = 5$.
2. *Population variance:* $\sigma^2 = 9$.
3. $E[\bar{x}] = \mu = 5$.
4. $var(\bar{x}) = \frac{\sigma^2}{n} = \frac{9}{5}$.
5. *Because each observation is assumed to follow the Normal distribution with the same parameters, then* $\bar{x} \sim N(E[\bar{x}], var(\bar{x})) \to \bar{x} \sim N(\mu, \frac{\sigma^2}{n}) = N(5, \frac{9}{5})$.
6. *If, however, it is only known that the observations are independent and are identically distributed, then, by the Central Limit Theorem,* $N(\bar{x}, \frac{s^2}{n})$ *will approximate the exact distribution of* $\bar{x}$, *and* $N(\bar{x}, \frac{s^2}{n})$ *will converge to* $N(\mu, \frac{\sigma^2}{n})$ *as the number of observations tend to infinity.*

Problem 4. *A random sample of size 36 elements is drawn from a population which the Normal, N(10,144), adequately describes.*

1. $E[\bar{x}] = \mu$.
2. $var(\bar{x}) = \frac{\sigma^2}{n} = \frac{144}{36} = 4$; *therefore,* $\bar{x} \sim N(10, 4)$.
3. $P(\bar{x} > 11) = P(\frac{\bar{x}-10}{\sqrt{4}} > \frac{11-10}{\sqrt{4}}) = P(z > \frac{1}{2}) = 1 - P(z \leqslant \frac{1}{2}) = 1 - (.5 + .1915) = .3085$.

Problem 5. *In a manufacturing plant, 50 products were weighted. Let* X_i *be the measurement of the weight of a product, i, in grams. Assume that the* X_i*'s are i.i.d. N(6,2.5) and that they adequately describe the measurements. Find the probability that the average weight will lie between 5.75 and 6.25 grams.*

1. $E[\bar{x}] = \mu = 6$.
2. $var(\bar{x}) = \frac{\sigma^2}{50} = \frac{2.5}{50} = .05$.
3. *In order to find,* $P(5.75 \leqslant \bar{x} \leqslant 6.25)$, *we first observe that* $\bar{x} \sim N(\mu, \frac{\sigma^2}{n}) = N(6, .05)$; *therefore,*

$$\begin{aligned}
P(5.75 \leqslant \bar{x} \leqslant 6.25) &= P\left(\frac{5.75-6}{\sqrt{.05}} \leqslant \frac{\bar{x}-6}{\sqrt{.05}} \leqslant \frac{6.25-6}{\sqrt{.05}}\right) \\
&= P(-1.118 \leqslant z \leqslant 1.118) \\
&= P(z \leqslant 1.118) - P(z \leqslant -1.118) \\
&= .5 + .3686 - (.1314) = .7372
\end{aligned}$$

Example 28. *$X_i \sim N(13, 9)$ for $X_1,\ X_2, \ldots, X_{36}$. Find $P(12.5 \leqslant \bar{X} \leqslant 13.5)$.*

The parameters of the sampling distribution are: $E[\bar{x}] = 13$ and $var(\bar{x}) = \frac{9}{36}$, so that $\bar{X} \sim N(13, \frac{9}{36})$ and:

$$\begin{aligned} P\left(\frac{12.5-13}{\sqrt{\frac{9}{36}}} \leqslant \frac{\bar{x}-13}{\sqrt{\frac{9}{36}}} \leqslant \frac{13.5-13}{\sqrt{\frac{9}{36}}}\right) &= P(-1 \leqslant z \leqslant 1) \\ &= P(z \leqslant 1) - P(z \leqslant -1) \\ &= .8413 - .1587 \\ &= .6826 \end{aligned}$$

Example 29. *Approximation of Binomial distribution by Normal distribution. Let $X \sim b(55, .373)$; find $P(X_i \leqslant 25)$. Use Central Limit Theorem where n=1 and that $E[X] = np = 55(.373) = 20.515$ and $var(X) = npq = 55(.373)(.627) = 12.8629$. Therefore computing the sample z-score within the probability statement gives: $P(\frac{X-20.515}{\sqrt{12.8629}} \leqslant \frac{25-20.515}{\sqrt{12.8629}}) = P(z \leqslant 1.2505) = .5 + .39435 = .8944$. The exact answer is: $P(X \leqslant 25) = .9164213\ldots$. So using the Normal distribution produces only an approximation for computing the probability of a Binomial distribution. Even in the case of one observation, the approximation is still rather good.*

Appendix A: Conditional Expectation and Conditional Variance

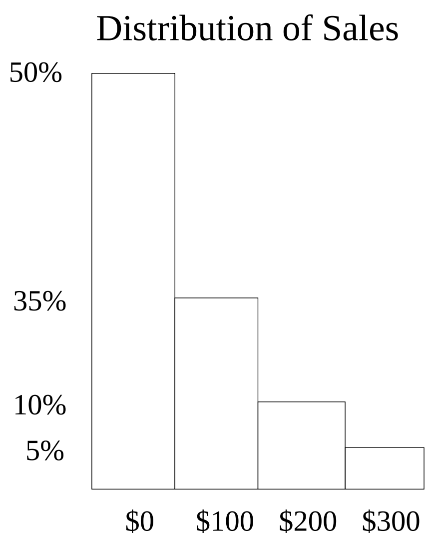


Figure 5

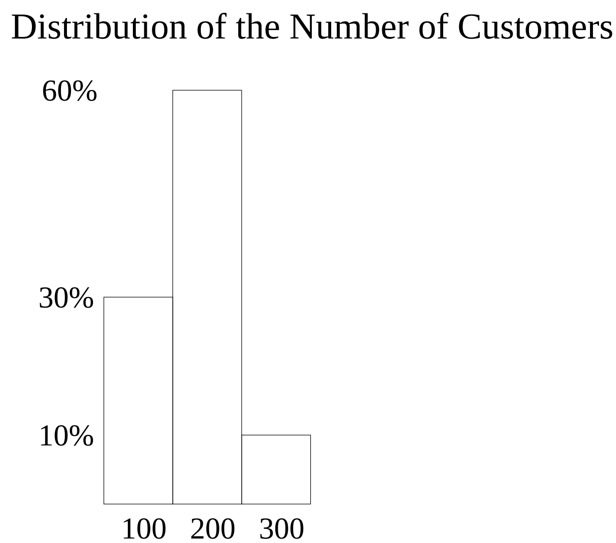


Figure 6

The owner of a store which sells perfume conducted a study of the number of customers which enter his store and the amount in sales which a customer makes.

$$S = \begin{cases} \$0 & 50\% \\ \$100 & 35\% \\ \$200 & 10\% \\ \$300 & 5\% \end{cases} \tag{2}$$

$$N = \begin{cases} 100 & 30\% \\ 200 & 60\% \\ 300 & 10\% \end{cases} \tag{3}$$

According to the owner's study, the distribution of sales is given in equation (2) and shown in Figure 5, and the distribution of the number of customers who enter the store is given in equation (3) and in Figure 6.

The owner wishes to find the expected sales on a day and the variance of the sales. To calculate those quantities, the owner will employ conditional expectation and conditional variance.

Theorem 13. *Given random variables, X and Y, then:*

$$\begin{aligned} E[Y] &= E[E[Y|X]] \quad (4) \\ var(Y) &= var(E[Y|X]) + E[var(Y|X)] \quad (5) \end{aligned}$$

Conditioning is utilized because the total sales is related to both the random variable N, the number of customers per day, and the random variable S, the amount of sales per customer. Let S_i be the amount which customer i purchases. Assume that customers will purchase perfume independently of another customer and that sales are identically distributed. Of course, a companion of a customer might recommend a certain brand of perfume and then purchase the same brand. To simplify the problem, however, we will assume that the customers act independently of each other.

We know from Definition 12 found on page 24 that the expected value of a random variable is:

$$E[X] = \sum_{\substack{all\ possible \\ values\ of\ X}} kP(X = k)$$

and from Definition 13 that the variance of a random variable is:

$$var(X) = \sum_{\substack{all\ possible \\ values\ of\ X}} (k - E[X])^2 P(X = k)$$

The expected value of S and of N as well as their variances are given below:

$$\begin{aligned} E[N] &= 100(.3) + 200(.6) + 300(.10) = 180 \\ var(N) &= (100 - 180)^2(.3) + (200 - 180)^2(.6) + (300 - 180)^2(.1) = 3600 \\ E[S] &= 0(.50) + 100(.35) + 200(.10) + 300(.05) = 70 \\ var(S) &= (0 - 70)^2(.50) + (100 - 70)^2(.35) + (200 - 70)^2(.10) + (300 - 70)^2(.05) = 7100 \end{aligned}$$

Define the total daily sales to be $T = \sum_{i=1}^{N} S_i$. The complication which is readily seen is that the upper limit of the sum is a random variable. If the upper limit was a constant, then we would be able to evaluate the summation. The use of conditional expectation will make N a constant for the purpose of evaluating the sum.

According to equation (4), we can write: $E[T] = E[E[T|N]]$. Given an N, then we find the expected value of T.

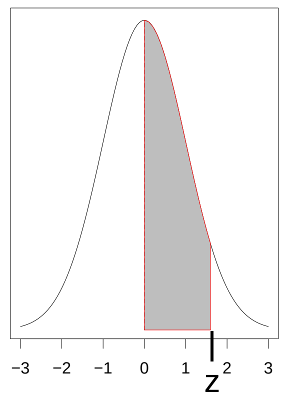
$E[T|N] = E[\sum_{i=1}^{N} S_i|N]$. By assumption, the S_i's are identically distributed, so that $E[S_i] = E[S] \quad \forall i$. Given N, $E[\sum_{i=1}^{N} S_i] = \sum_{i=1}^{N} E[S] = NE[S]$. Therefore, $E[T|N] = NE[S]$, and $E[T] = E[E[T|N]] = E[NE[S]] = E[N]E[S] = 180(70) = \$12,600$. This agrees with our intuition; the total sales is the product of the average sale per customer by the expected number of customers. The calculation of the variance of the total sales, on the other hand, is as always the challenge. Equation (5) will make the calculation possible.

In the context of the owner's perfume business, $var(T) = var(E[T|N]) + E[var(T|N)]$. For the first term, we already calculated $E[T|N] = NE[S]$; therefore, the first term of equation (5) is $var(E[T|N]) = var(NE[S]) = var(N)E[S]^2$, because, according to Theorem 9 found on page 52, $E[S]$ being a constant comes out of the variance as a square. We complete the calculation as $var(E[T|N]) = var(N)E[S]^2 = 3600(70^2) = 17640000$

We, now, need to address the second term of equation (5), $E[var(T|N)]$. To begin with, $var(T|N) = var(\sum_{i=1}^{N} S_i|N)$. By assumption, S_i's are independently and identically distributed. Given N, $var(\sum_{i=1}^{N} S_i) = \sum_{i=1}^{N} var(S) = Nvar(S)$; therefore, $E[var(T|N)] = E[Nvar(S)] = E[N]var(S)$ because $var(S)$ is a constant and by referring again to Theorem 9 we are led to $E[var(T|N)] = E[N]var(S) = 180(7100) = 1278000$.

By combining both terms, we get $var(T) = 17640000 + 1278000 = 18918000$ and its square root is: 4349. Based on the owner's study, the expected sales is $\$12,600 \pm 4349$ with a CV of $\frac{4349}{12600} = 34.5\%$.

When this problem is view from the perspective of quality control, we observe that the first term of $var(T)$ is 14 times larger than the second term. In the first term, $var(E[T|N]) = var(N)E[S]^2$, the variablity in the number of customers, $var(N)$, causes the very large value of the first term. In other words, the closer the variance of N is to zero, the more certain the owner of the perfume store will be in estimating the amount of daily sales. He might want to launch a new advertising campaign in such as way as to create a more uniform flow of customers and, at the same time, possibly to motivate the window shoppers to purchase even an inexpensive perfume, in order to drive down $var(S)$.



Cumulative Probabilities for a N(0,1) Distribution: $\Phi(z) - .5$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.0279	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41309	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2.0	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.49900