Prof. Sergio A. Alvarez
21 Campanella Way, room 569
Computer Science Department
Boston College
Chestnut Hill, MA 02467 USA

http://www.cs.bc.edu/∼alvarez/
alvarez@cs.bc.edu
voice: (617) 552-4333
fax: (617) 552-6790

CS244, Randomness and Computation
Probability Axioms and Basics

Axiomatic Probability

Probability was developed through centuries of work motivated by problems such as the odds in a card game and the large-scale molecular dynamics of gases. From this development, there arose a set of basic notions upon which probability calculations, in any context, can be based. The behavior of these basic objects is described by a set of *axioms*, as we call them, which are statements accepted without question. This is very much like the axioms of Euclid's geometry, which introduce basic notions such as point, line, and their fundamental properties (e.g., if point P is not on line L, there is a unique line L' that passes through P and does not intersect L). These notes present the axioms of probability theory and some of their basic consequences. We also consider issues of counting that are important in performing calculations involving the axioms, in the important special case of equally likely outcomes.

Discrete probability axioms

As described in the Introduction, probability deals with random experiments that can be repeated many times. The axioms of probability formalize the basic ingredients that are needed to specify such a random experiment in order to allow probability calculations related to it.

We begin with discrete probability, which applies to experiments that yield outcomes belonging to a set that is either finite or whose elements can be enumerated in a list indexed by the counting numbers $1, 2, 3, \cdots$. This eliminates sets that are "continuous", such as intervals of real numbers. We will come back to the latter case soon enough. Without further ado, I now present our basic cast of characters in the case of discrete probability.

- The *sample space*: a set S consisting of all possible outcomes of a random experiment. In the discrete case, this set is assumed to be finite or countable as described above.
- The *probability function*, also known as the *probability mass function* or *probability distribution function*, $p : S \to [0, 1]$, that assigns a probability value to each possible outcome in the sample space.

The combination of a sample space together with a probability function is sometimes called a *probability space*, perhaps because it sounds fancier that way.

Comments on the basic elements. Axioms of discrete probability.

1. The sample space S is completely unconstrained except for the temporary requirement of discreteness, which we shall eliminate later. It can comprise the results of a coin toss, or a hand of cards that has been dealt, or the momentum resulting from a molecular collision (assuming for now that the momentum is expressed in finite precision arithmetic).

2. The probability function is required to satisfy the following properties:

   - p takes real values between 0 and 1.
   - The sum of all of the values $p(x)$, for x in the sample space, equals 1.

   The interpretation of the probability values is that the probability of an outcome is a measure of the relative frequency with which that particular outcome is observed in the random experiment being considered. For example, $p(\text{green light}) = 1/2$ would mean that a green light is expected in one half of all instances of the experiment in question (whatever the context happens to be).

Examples of instances of the discrete probability axioms.

Tossing a fair coin. If there is such a thing as *the* most classic example of a probability space, this is it. The sample space consists of the two possible outcomes, heads and tails: $S = \{\text{heads}, \text{tails}\}$. The probability function simply states that heads and tails are equally likely to occur: $p(\text{heads}) = 1/2$, $p(\text{tails}) = 1/2$.

Hitting the bull's eye on a dartboard. The sample space consists of two possible outcomes: hit, and miss. The details of the probability function depend on who is throwing the darts. *We're just constructing a model here.* Someone who is very good at the game may be described by $p(\text{hit}) = 0.9$, $p(\text{miss}) = 0.1$, while your nearsighted cousin who refuses to get eyeglasses may be closer to: $p(\text{hit}) = 0.2$, $p(\text{miss}) = 0.8$.

Bullseye, take 2. Notice that one could also model the dart-throwing experiment in an alternative way, if differentiating among the non-bullseye regions of the target is important. The sample space S could correspond to the various regions that can be hit. For example, suppose that the dartboard is divided by concentric circles into regions with point values of 100, 80, 60, 40, and 20, with 100 being a bullseye. The sample space can then be made to have five elements, each corresponding to one of the different regions. The probability function would then be given by $p(\text{simple region}) = 0.2$, where the simple region is any one of the five in the sample space $S = \{100, 80, 60, 40, 20\}$.

Dealing two cards from a deck. A standard deck of playing cards contains 52 cards, with the 13 cards Ace, 2, 3, 4, 5, 6, 7, 8, 9, Jack, Queen, King for each of the four suits Hearts, Diamonds, Spades, and Clubs. Suppose that a particular card game involves dealing two cards at random from a full deck. You can model the experiment of dealing two cards in terms of the sample space that consists of all possible two-card hands that can be dealt: $S = \{\{c, c'\} | c$ and c' are different cards$\}$. Notice that it does not matter what order the cards are dealt in, so each outcome is an *unordered set* rather than an ordered pair. Small detail, but a significant one. We assume that all pairs are equally likely, since the cards are dealt at random. We model this assumption with the uniform probability function that assigns the same probability value to all hands. What value is this? Well, since the probabilities of all outcomes in S must sum to 1, this common probability value must equal 1 divided by the total number of outcomes in S. By counting, you can figure out that there are exactly 1326 such outcomes. Therefore, $p(\{c, c'\}) = 1/1326$ for all hands $\{c, c'\}$.

Customers or processes arriving at a server. Queuing theory models the arrival and FIFO (first-in first-out) service of requests at a server. An often used model of arrivals per unit time is specified by the infinite (but discrete) sample space $S = \{0, 1, 2, 3, \cdots n \cdots\}$ of all non-negative integers, together with the probability function $p(n) = e^{-r} r^n / n!$ that measures the probability that exactly n requests will arrive within one unit of time. The value r is a parameter that represents the average number of arrivals per unit time. The fact that this function $p(n)$ satisfies the probability axioms follows from convergence of the power series expansion of the exponential function: $\sum_{n=0}^{\infty} r^n/n! = e^r$, together with the fact that $e^{-r}e^r = 1$. This probability function is illustrated in Fig. 1 in the case $r = 3$.

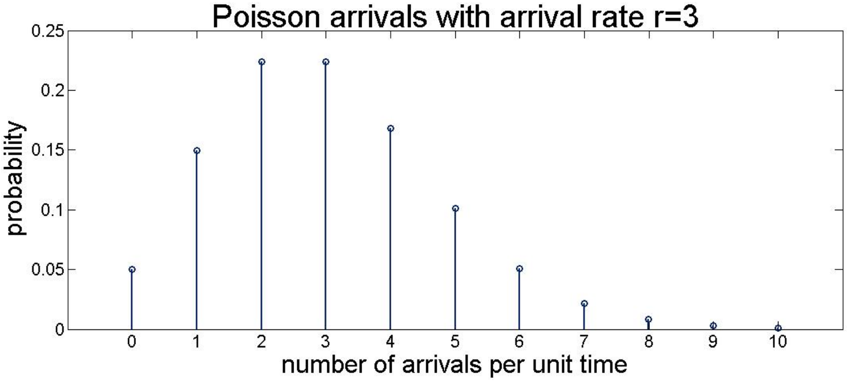

Figure 1: Probability of n arrivals per unit time in a Poisson model.

Power-law distributions. The Poisson probability distribution considered in the service request arrivals model above decays exponentially fast with the number of service requests. In contrast with this behavior, many phenomena are better modeled with a probability function that decays slowly on a sample space of integers. The best known of these are the power-law distributions $p(n) = an^{-b}$ on the set of positive integers $n = 1, 2, 3, \cdots$, where the exponent b is a real-valued constant greater than 1 and the factor a is a normalization constant that guarantees that the values $p(n)$ sum to 1. Power-law distributions have been fit empirically as models of the distribution of file sizes in computer systems (n is the file size), or of word frequency in a natural language (n is the popularity rank of a word). See Fig. 2 for an example with exponent $b = -1.1$.

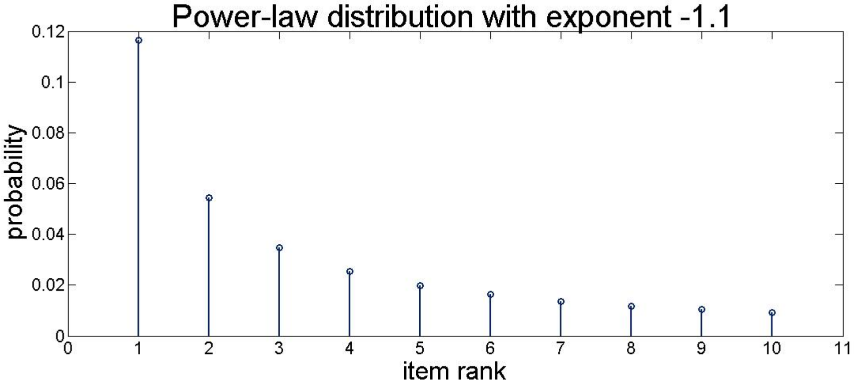

Figure 2: Probability of items as a function of rank in a power-law model.

Random events

A probability space (a sample space together with a probability function on the outcomes in the sample space) specifies the probability of every possible outcome of the random experiment being modeled. Often, however, one is interested in determining the probability of a set of outcomes all at once. For example, if the experiment consists of rolling a six-sided die once, then the elementary outcomes are just the face values 1,2,3,4,5,6, while one could be interested in the probability of rolling an even number as opposed to an odd number. This composite outcome can be viewed as the set $\{2, 4, 6\}$ that contains all possible even elementary outcomes. Another example is the two-card hand experiment described above. In that case, one could be interested in knowing the probability that the hand will contain an Ace. Again, the object of interest here is not one of the elementary outcomes (a set of two specific cards), but rather a class containing many of these outcomes, namely the set of all two-card hands consisting of an Ace and some other card (perhaps another Ace also, perhaps not). Composite outcomes such as these are called *events*.

Extension of the probability function to random events.

A probability function on the sample space provides a straightforward way to assign probabilities to events in terms of their constituent elementary outcomes. An event occurs if and only if one of its constituent elementary outcomes occurs. Therefore, *in order to maintain* *the interpretation of probability as expected frequency of occurrence, the probability of an event should be the sum of the probabilities of the elementary outcomes that comprise it.* We will *define* the probability of an event in this way.

Properties of the probability function on events. The extended function on random events, denoted by P, has several properties that we now discuss. For the moment, we consider these properties to be consequences of the above definition of probability of an event as the sum of the probabilities of the constituent elementary outcomes. Later, in order to accomodate non-discrete sample spaces, we shall recast the foundations of probability in terms of events, by making these properties new axioms.

- *Non-negativity of probability:* The probability of an event is non-negative:

  $$P(E) \geq 0 \text{ for all events } E$$

  This follows from the axiom for non-negativity of the probability function p on elementary outcomes.

- *Probability of the sample space:* The probability of the event defined by the full sample space is 1:

  $$P(S) = 1$$

  Given the definition of probability of an event, this is just a restatement of the axiom that requires that the values of the probability function p on elementary outcomes sum to 1.

- *Additivity:* The probability of a disjoint enumerable union of events is the sum of the probabilities of the events forming the union:

  $$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) \text{ if } A_i \cap A_j = \emptyset \text{ whenever } i \neq j$$

  This follows from the definition of probability of an event, as the sum that defines the probability of A may be expressed as the sum of individual sums corresponding to the probabilities of the events A_i.

The following is a consequence of the above fundamental properties:

- *Complementary probability:* The probability of the event consisting of all elementary outcomes that do *not* belong to an event is 1 minus the probability of that event:

  $$P(S \setminus A) = 1 - P(A)$$

  This follows from the properties of additivity and probability of the sample space for the extended probability function P described above.

Examples of random events and calculation of their probabilities.

Probability of the Queen of Hearts in a two-card hand. In the two-card hand example above, what is the probability that a hand will contain the Queen of Hearts? Well, it's just the sum of the probabilities of all hands that contain the Queen of Hearts. Since all hands have the same probability, this is just 1/1326 multiplied by the total number of hands that contain the Queen of Hearts. Any such hand consists of the Queen of Hearts and exactly one other card. There are precisely 51 other possible cards (all other cards in the deck). Therefore, the probability of being dealt the Queen of Hearts in a two-card hand is 51/1326.

Probability of a prime number in a random roll of a fair six-sided die. The fair die can be modeled by the sample space $S = \{1, 2, 3, 4, 5, 6\}$ together with the uniform probability function $p(i) = 1/6$ for each value of i between 1 and 6. The event of rolling a prime number is the set consisting of all elementary outcomes that happen to be prime numbers: $\{2, 3, 5\}$. The probability of this event is the sum of the elementary outcomes that belong to it, which is $1/6 + 1/6 + 1/6 = 1/2$.

Probability of at least one Tails in two flips of a fair coin. We employ the sample space $S = \{$(Heads,Heads), (Heads,Tails), (Tails,Heads), (Tails,Tails) $\}$ of all ordered pairs of faces, with the probability function $p(f_1, f_2) = 1/4$ for each elementary outcome. The target event A is the complement of the event $\{$(Heads,Heads)$\}$. Hence, the desired probability is $1 - 1/4 = 3/4$.

Calculating event probability for equiprobable outcomes. Counting.

As shown in the above discussion, if all elementary outcomes have the same probability then the probability of any event is just the probability of an elementary outcome multiplied by the number of elementary outcomes that contribute (belong) to the event. Therefore, the problem of calculating the probability reduces to the problem of counting the number of elements in the event in the case of equiprobable elementary outcomes. We now briefly address techniques for solving such counting problems. Throughout, the symbol $|A|$ denotes the number of elements in the set A.

Fundamental principles of counting.

The partition principle. If set A is the union of nonoverlapping (pairwise disjoint) sets $A_1, \cdots A_n$, then the number of elements of A is the sum of the numbers of elements in the sets A_i:

$$|A| = \sum_{i=1}^{n} |A_i| \text{ if } A = \bigcup_{i=1}^{n} A_i \text{ and } A_i \cap A_j = \emptyset \text{ whenever } i \neq j$$

The multiplication principle. If set A is the set of all ordered sequences $a_1 \cdots a_n$ of elements of sets $A_1 \cdots A_n$, then the number of elements of A is the product of the numbers of elements of the sets A_i:

$$|A| = \prod_{i=1}^{n} |A_i| \text{ if } A = \{(a_1 \cdots a_n) | a_i \text{ belongs to } A_i \text{ for each } i = 1 \cdots n\}$$

Examples of counting using the fundamental counting principles.

Number of five-letter pretend-words that start with “gate”. A pretend-word is just a sequence of letters from the English alphabet a–z. The five-letter pretend words that start with “gate” form a set A that may be expressed as the disjoint union of the 26 sets $A_a \cdots A_z$, where, for any letter x in the alphabet a-z, A_x consists of all five-letter pretend-words that start with “gate” and end with the letter x. Each of these sets A_x contains exactly one word. Therefore, by the addition principle, A contains exactly 26 words.

Number of possible ordered outcomes in three rolls of a four-sided die. Each outcome is an ordered triple of the form (f_1, f_2, f_3), where each f_i is one of the face values 1,2,3,4. By the multiplication principle, there are exactly $\prod_{i=1}^{3} 4 = 4^3 = 64$ such outcomes.

Number of possible two-card hands, both Hearts, drawn from the same deck. Cards of the same suit are differentiated by their face value, of which there are exactly 13 varieties (Ace, 2, 3, 4, 5, 6, 7, 8, 9, Jack, Queen, King). Hands are *not* ordered pairs, but rather unordered sets. However, notice that if we can count all ordered pairs of Hearts, we can just divide by 2 at the end to account for the fact that there are exactly two ordered pairs for every hand, one for each of the two possible orders in which the two cards of the pair can be dealt. For example, we count the two ordered sequences $(\text{Hearts}, 2)$ and $(2, \text{Hearts})$ as the same hand. In order to count ordered pairs, notice that the set A of all ordered pairs of face values can be expressed as the disjoint union of thirteen sets A_i, where A_i contains all pairs for which the first card has the i-th possible face value. By the addition principle, the number of elements of A is the sum of the numbers of elements of all of the A_i. Each A_i, in turn, consists of all ordered pairs of Hearts in which the first component has the i-the face value; since both cards are drawn from the same deck, only 12, not 13, face values are possible for the second component of each pair in A_i. Therefore, each A_i contains exactly 12 elements. We conclude that A contains $\sum_{i=1}^{1} 3|A_i| = \sum_{i=1}^{1} 312 = 156$ elements. By the ordered-to-unordered conversion discussed above, the total number of two-card Hearts hands is $156/2 = 78$.

Number of two-card hands in which both cards are of the same suit. Here, the set A consists of all two-card hands $\{c, c'\}$ for which c and c' have the same suit. A may be expressed as the disjoint union of the sets A_i, $i = 1...4$, where A_1 consists of all hands of two Hearts, A_2 consists of all hands of two Diamonds, A_3 consists of all hands of two Spades, and

A_4 consists of all hands of two Clubs. Since each A_i contains exactly 78 elements (see the preceding example), the set A contains exactly 4 times $78 = 312$ elements, by the addition principle.

Probability Axioms in the General Case

Above, we discussed an axiomatic foundation for discrete probability based on the notion of sample space and of a probability distribution function defined on the elementary outcomes in the sample space. We would like to be able to address continuous sample spaces as well. However, assigning probability values to the elementary outcomes of a continuous sample space is doomed to fail, as the following example shows.

Example: timing with infinite precision. Suppose that a physical experiment produces a detectable output, say turning on the light, at a completely random time between 0 and 1 seconds after starting the experiment. What is the probability that the light will turn on at exactly 0.5 seconds? We assume that times have infinite precision, so that 0.5 seconds is entirely different than 0.4999999 seconds or 0.500001 seconds, for example (and those two examples just use a few digits after the decimal point). Since probability is interpreted as relative frequency, this question amounts to the question of what fraction of repetitions of this experiment can be expected to lead to the light turning on at *precisely* 0.5 seconds. The inevitable answer is zero! In many repetitions, you may observe a few times between 0.499 and 0.501. In fact, there is no way to entirely rule out the near impossibility of observing a time of precisely 0.5 seconds. However, you most definitely cannot *expect* to observe a time of exactly 0.5 seconds in any nonzero fraction of the experimental runs.

Individual elementary outcomes in a continuous sample space, for example the points in an interval of real numbers as in the above example, will usually have zero probability. On the other hand, it is clear that many events of interest, such as the time falling between 0.5 and 0.75 seconds, can be expected to be observed in a nonzero fraction of repetitions of the random experiment and therefore should have a corresponding nonzero probability. Because of this, the simple rule that equates the probability of an event with the sum of the probabilities of its constituent elementary outcomes, does not hold for general sample spaces (the sum of many zeros is still zero). This fact forces a recasting of the foundations of probability. Fortunately, the solution is not complicated: we merely have to place *events*, rather than elementary outcomes, at the center.

Measurable events. General probability axioms. A probability space, which is meant to model a random experiment, consists of the following components:

- A set S, called the *sample space*, the elements of which are the *elementary outcomes* of the random experiment

- A collection of subsets of S called *measurable events* that has the following properties[1]:
  - The sample space S is a measurable event
  - The complement of any measurable event is a measurable event
  - The union of any (finite or infinite) enumerable family $S_i,\ i = 1, 2, 3, ...,$ where each S_i is a measurable event, is a measurable event
- A probability function P defined on the set of measurable events, and that has the following properties:
  - *Non-negativity of probability:* The probability of an event is non-negative:

    $$P(E) \geq 0 \text{ for all events } E$$
  - *Probability of the sample space:* The probability of the event defined by the full sample space is 1:

    $$P(S) = 1$$
  - *Additivity:* The probability of a disjoint enumerable union of events is the sum of the probabilities of the events forming the union:

    $$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) \text{ if } A_i \cap A_j = \emptyset \text{ whenever } i \neq j$$

Comments on the general probability axioms. In our discussion of discrete probability, we derived the properties expressed by the above general axioms as direct consequences of defining the probability of an event as the sum of the probabilities of the elementary outcomes that comprise that event. In the general case, it is no longer possible to assign elementary outcome probabilities in a useable way, as shown by the infinite precision timing example. Hence, we simply make these properties of event probability the new *definition.* Nice trick, huh? We know that doing this will not create any inconsistencies with the old definition of discrete probability, since the new definition follows from the old one. Furthermore, we will now be able to deal with more general sample spaces, as well as with discrete sample spaces, within a single framework.

Examples of instances of the general probability axioms.

Infinite precision timing. Let's try this again. We wish to model an experiment that produces a random real-valued time between 0 and 1. We will construct a probability space that provides such a model.

[1] For most of our discussions, you can ignore the technical requirements on the set of measurable events, and simply equate events with arbitrary subsets of the sample space.

- The sample space is easy: $S = [0, 1]$, the closed unit interval, consisting of all real numbers between 0 and 1, which are all of the possible elementary outcomes of the experiment.

- What are the measurable events? This is trickier. It seems clear that any finite precision measurement should be allowed as a special case, and that there will be a measurable (and nonzero) probability that the time will fall within a *subinterval* of real numbers, corresponding to such coarse-grained, or finite precision, timing. Therefore, all subintervals $[a, b]$, where $0 \leq a \leq b \leq 1$, should be measurable events, so that a well-defined probability value will be assigned to them. Finite and even enumerably infinite unions of subintervals will then also automatically be granted measurable event status, by the axioms for the set of measurable events. Complements of such unions are unions of the same kind, as long as open intervals are allowed together with closed intervals. Therefore, we have a reasonable set of measurable events: all enumerable unions of subintervals of $[0, 1]$.

- The probability function should assign to each event a measure of the expected relative frequency with which times are observed within that event (set). Since the measurable events here are enumerable unions of intervals, and since the probability of such a union is just the sum of the probabilities of the individual terms of the union, it is enough to define the probability of a subinterval. The additivity axiom will then extend the probability values to all measurable events. So, what should the probability $P([a, b])$ be? A "random spraying" of times between 0 and 1 will hit the interval $[a, b]$ with about the same relative frequency as the ratio of the length of $[a, b]$ to the length of $[0, 1]$. This means that we should let $P([a, b]) = b - a$. For example, the interval $[0, 1/2]$ has probability $1/2$ since it occupies one half of the sample space, while the interval $[0.7, 0.8]$ has probability 0.1.

We can test the above ideas by simulating the experiment in MATLAB. The rand function generates uniformly distributed pseudorandom numbers between 0 and 1, and provides a model for the times described in the above random times example. I generated a random vector of $10,000$ times using the rand function, and tabulated the observed frequency with which these times fall within the intervals $[0.1(k-1), 0.1k]$ for $k = 1...10$. See Fig. 3. The results show approximately $1,000$ occurrences, or 10% of the total number of observations, within each of these subintervals. This is consistent with the assignment of the probability value 0.1 to each of them.

Exponentially distributed arrivals. Consider the times between consecutive arrivals of service requests at a server. The model of random arrival times described below is, as we shall better understand later, very natural. Unlike the uniform distribution model considered above, in which there is a finite interval within which all observed times occur, the model below allows arrivals to occur at any non-negative time. However, the likelihood that an arrival will occur near a given time is highest for small times and decreases rapidly with increasing time.

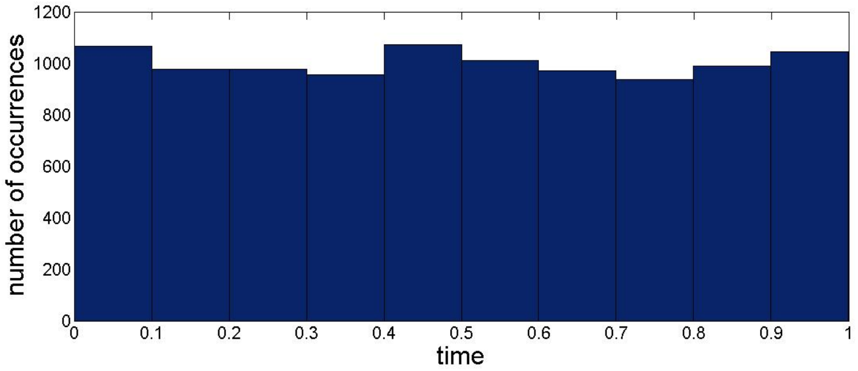

Figure 3: Number of occurrences of random times between 0 and 1 in $10,000$ observations.

- The sample space is the set of non-negative *real* numbers, as we impose no predetermined time horizon for the next arrival.

- Measurable events will be taken to be unions of intervals as in the timing example.

- The probability function is defined in an interesting way: by specifying its "infinitesimal" behavior, that is, through calculus. The probability that the next arrival will be observed in the length dt interval $[t - dt/2, t + dt/2]$ is $e^{-t}dt$ if dt is "infinitesimally small".[2] The function e^{-t} here plays the role of a *probability density.* It quantifies the likelihood of observing times in a small interval near a particular point t as a factor times the length of the interval. For a given length, the likelihood is highest for small values of t, and decreases very rapidly as t gets larger. Thus, the first arrival is expected to occur within a few time units in most cases. The probability of the measurable event $[a, b]$ is the sum of the probabilities of the infinitesimal intervals that span $a, b]$:

$$P([a, b]) = \int_a^b e^{-t}\, dt$$

  This probability assignment satisfies all of the general probability axioms. It is interesting to note that this model of the times between consecutive arrivals implies that the *total number* of arrivals in a given time interval satisfies the Poisson model described in an earlier example.

A simulation of the exponential arrivals model described above produces the histogram of observed frequencies shown in Fig. 4, in which 100 bins are used. The observed distribution of times very closely follows the graph of the exponential function that defines the probability assignment in this model.

[2] If you are concerned about foundational difficulties associated with the definition of infinitesimal, relax. We will only be using this language informally in order to motivate the subsequent integrals.

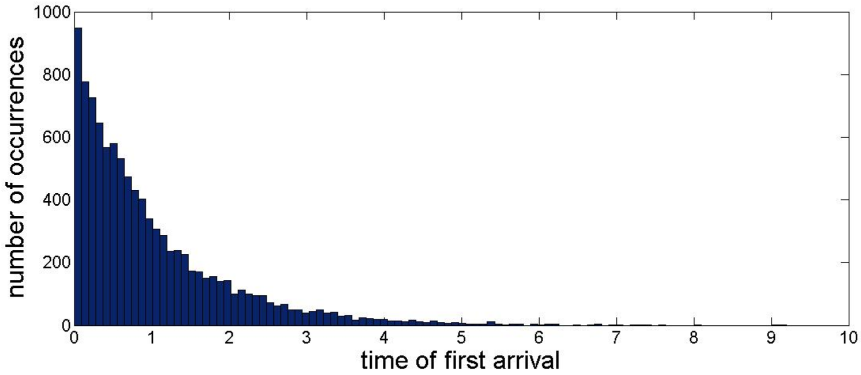

Figure 4: Histogram of exponentially distributed non-negative times in $10,000$ observations.

Exercises.

1. For each of the following, give a sample space and probability function that provide a good framework in which to analyze the situation described. You need not calculate the probability of the event described.

   (a) Rolling a total of 6 on three rolls of a fair 4-sided die.

   (b) Sequentially drawing two cards in increasing face value from a standard playing deck of 52 cards.

2. Explain how to compute the value of the normalization constant a that is needed in order for the function $p(n) = an^{-b}$, where $b > 1$, to be a valid probability function on the set of all positive integers $n = 1, 2, 3, \cdots$. Is the restriction $b > 1$ really necessary here? Explain.

3. Working from basic principles, count the number of different possible two-card hands that can be drawn from a single standard playing deck of 52 cards.

4. Calculate the probability of each of the following random events.

   (a) Getting two Hearts in a random two-card hand from a standard deck.

   (b) Getting cards of two different suits in a random two-card hand from a standard deck.

   (c) Rolling strictly increasing values in three consecutive rolls of a fair four-sided die.

5. Show in detail that the function $P([a, b]) = \int_a^b e^{-t}\, dt$ satisfies general probability axioms on the set of subintervals of non-negative real numbers.