Theoretical Probability and Statistics

Charles J. Geyer

September 5, 2008

Contents

Chapter 1

Finite Probability Models

The fundamental idea in probability theory is a *probability model*, also called a *probability distribution*. Probability models can be specified in several different ways

- probability mass function (PMF),
- probability density function (PDF),
- distribution function (DF),
- probability measure, and
- function mapping from one probability model to another.

We will meet probability mass functions and probability measures in this chapter, and will meet others later. The terms "probability model" and "probability distribution" don't indicate how the model is specified, just that it is specified somehow.

1.1 Probability Mass Functions

A *probability mass function* (PMF) is a function

$$\Omega \xrightarrow{\text{pr}} \mathbb{R}$$

which satisfies the following conditions: its values are nonnegative

$$\text{pr}(\omega) \geq 0, \qquad \omega \in \Omega \tag{1.1a}$$

and sum to one

$$\sum_{\omega \in \Omega} \text{pr}(\omega) = 1. \tag{1.1b}$$

The domain Ω of the PMF is called the *sample space* of the probability model. The sample space is required to be nonempty in order that (1.1b) make sense.

In this chapter we require all sample spaces to be finite sets. Elements of the sample space are called *outcomes*. and values of the PMF are called *probabilities*: $\operatorname{pr}(\omega)$ is the *probability* of the outcome ω.

That's all there is to it. All of probability theory is built on this one idea and generalizations of it.

> *Probability theory is the study of functions that are nonnegative and sum (or integrate) to one.*

There is nothing special about the notation. We could just as well specify a model by a finite set S and a real-valued function f satisfying

$$f(x) \geq 0, \qquad x \in S$$

and

$$\sum_{x \in S} f(x) = 1.$$

Or we could just as well specify a model by a finite set $\{x_1, \ldots, x_n\}$ and a real-valued function g satisfying

$$g(x_i) \geq 0, \qquad i = 1, \ldots, n$$

and

$$\sum_{i=1}^{n} g(x_i) = 1.$$

The fundamental idea of a function that is nonnegative and sums to one can be expressed in many different notations, but the underlying concept remains the same.

> *Mathematics is invariant under changes of notation.*

A PMF is a function whose values are nonnegative and sum to one. This concept can be expressed in many different notations, but the underlying concept is always the same.

Learn the concept not the notation. We will use all of the notations above and more for PMF. You must learn to recognize the concept clothed in any notation.

Theorem 1.1. *For any probability mass function* pr *on a sample space* Ω

$$0 \leq \operatorname{pr}(\omega) \leq 1, \qquad \omega \in \Omega.$$

Proof. That $\operatorname{pr}(\omega) \geq 0$ for all ω is a defining property (1.1a). That $\operatorname{pr}(\omega) \leq 1$ for all ω follows immediately from (1.1b) and (1.1a). □

Example 1.1.1 (One-Point Sample Space)**.**
A sample space cannot be empty. The smallest possible has one point, say $\Omega = \{\omega\}$. Then $\operatorname{pr}(\omega) = 1$.

This probability model is of no interest in applications. It is merely the simplest of all probability models. Compare with Example 1.1.5.

Example 1.1.2 (Two-Point Sample Space)**.**
The next simplest possible probability model has a sample space with two points, say $\Omega = \{x_1, x_2\}$. In this case, say $\text{pr}(x_1) = p$. Then we know from Theorem 1.1 that $0 \le p \le 1$. Now from (1.1b) we derive $\text{pr}(x_2) = 1 - p$.

The function pr is determined by one real number p

$$\text{pr}(\omega) = \begin{cases} p, & \omega = x_1 \\ 1 - p, & \omega = x_2 \end{cases}$$

For each different value of p, we get a different probability model. For example, $p = 1/3$ implies $\text{pr}(x_1) = 1/3$ and $\text{pr}(x_2) = 2/3$.

Definition 1.1.1 (Bernoulli Distribution)**.**
Any probability distribution on the sample space $\{0, 1\}$ *is called a* Bernoulli distribution. *If* $\text{pr}(1) = p$*, then we use the abbreviation* $\text{Ber}(p)$ *to denote this distribution.*

The Bernoulli distribution is our first "brand name" distribution.

REVISED DOWN TO HERE

In the preceding example it doesn't matter much what the points x_1 and x_2 are. Ω could be any two-point set and the probabilities of the probability model would be much the same. Therefore we usually use the special two-point set $\{0, 1\}$. When we are really interested in some other two-point set, say $\{\text{apple}, \text{orange}\}$, we say we have coded the outcomes, identifying perhaps apple with zero and orange with one. The point of this will become clear as we go along.

In Example 1.1.2 and Definition 1.1.1 we first meet the important concept of a family of probability distributions, which we also call a *statistical model* — a statistical model is a family of probability models. We often say, in a rather sloppy use of terminology, the "Bernoulli distribution" when we really mean the Bernoulli family of distributions, the set of all $\text{Ber}(p)$ distributions for $0 \le p \le 1$. As always when we use the word "distribution" or "model" we are not indicating exactly how the distribution or distributions are specified. It might be by probability mass functions. Define

$$f_p(x) = \begin{cases} 1 - p, & x = 0 \\ p, & x = 1 \end{cases}$$

Then whenever $0 \le p \le 1$ the function f_p is the PMF of a $\text{Ber}(p)$ distribution. We can think of the Bernoulli statistical model as this family of PMF's

$$\{\, f_p : 0 \le p \le 1 \,\}.$$

In the expression $f_p(x)$ both p and x are mathematical variables, but they play different roles, which are indicated by the notation. Each choice of a value

for p gives a function f_p that maps values x in the domain of the function, which in this case is the sample space $\{0, 1\}$, to values $f_p(x)$ in the range of the function, which for a PMF is always the closed interval $[0, 1]$.

Returning to our notation for Bernoulli PMF, f_p is a function and $f_p(x)$ is its value at the point x. We also say x is the argument of the function f_p. This allows us to say that p is not the argument of the function f_p. We need another term for it, and the established term is "parameter." A *parameter* is a real number that specifies a particular probability distribution within a family. For example, p is the parameter of the Bernoulli family of distributions. Each value of p in the range $0 \le p \le 1$ specifies a particular distribution $\text{Ber}(p)$.

Example 1.1.3 (Three-Point Sample Space)**.**
The next simplest possible probability model has a sample space with three points, say $\Omega = \{x_1, x_2, x_3\}$. In this case, say $\text{pr}(x_1) = p_1$ and $\text{pr}(x_2) = p_2$. Now from (1.1b) we derive $\text{pr}(x_3) = 1 - p_1 - p_2$.

The function pr is determined by two real numbers p_1 and p_2

$$\text{pr}(\omega) = \begin{cases} p_1, & \omega = x_1 \\ p_2, & \omega = x_2 \\ 1 - p_1 - p_2, & \omega = x_3 \end{cases}$$

Example 1.1.3 describes a family of distributions with two parameters p_1 and p_2. We sometimes collect multiple parameters into a single object, which is a vector, in this case the vector $\mathbf{p} = (p_1, p_2)$. Then we could write the PMF of this distribution either as f_{p_1,p_2} or the simpler $f_{\mathbf{p}}$.

The set of allowed parameter values, those that correspond to a distribution, is called the *parameter space* of a statistical model For Example 1.1.3 the parameter space is

$$\{\, (p_1, p_2) \in \mathbb{R}^2 : p_1 \ge 0 \text{ and } p_2 \ge 0 \text{ and } p_1 + p_2 \le 1 \,\}$$

where $\mathbb{R}^2$ is the Euclidean plane, the space of all two-dimensional vectors.

Example 1.1.4 (Discrete Uniform Distribution)**.**
Let $\{x_1, \ldots, x_n\}$ denote the sample space. The word "uniform" means all outcomes have equal probability, in which case (1.1b) implies

$$f(x_i) = \frac{1}{n}, \qquad i = 1, \ldots, n$$

defines the PMF. Later we will meet another uniform distribution, the continuous uniform distribution. The word "discrete" is to distinguish this one from that one.

Example 1.1.5 (Distribution Concentrated at One Point)**.**
Let S denote the sample space, and let x^* be a point in S. The distribution with PMF defined by

$$f(x) = \begin{cases} 1, & x = x^* \\ 0, & x \ne x^* \end{cases}$$

is called the distribution concentrated at x^*.

More generally, if S is the sample space of a probability distribution and f is the PMF, then we say the *support* of this distribution is the set

$$\{\, x \in S : f(x) > 0 \,\},$$

that is, $f(x) = 0$ except for x in the support.

In finite probability models outcomes having probability zero are unimportant in applications, so it makes little practical difference if such points are deleted from the sample space (so the new sample space is the old support). However, we often do not delete points of probability zero for reasons of mathematical convenience. $\text{Ber}(p)$ distributions for $0 < p < 1$ have support that is the same as the sample space $\{0, 1\}$, but the two remaining Bernoulli distributions with parameter values $p = 0$ and $p = 1$ are each concentrated at one point, 0 and 1, respectively. For reasons of convenience, we say all Bernoulli distributions have the same sample space, but two of them have different support.

None of the examples discussed so far are particularly important in themselves. Even the the only "brand-name distribution" we have discussed, the Bernoulli distribution, is important as a building block for other distributions. We need to learn some more theory before we can handle more interesting distributions. The examples discussed so far will have to serve for now.

1.2 Events and Measures

A subset of the sample space is called an *event*, and the probability of an event A is defined by

$$\Pr(A) = \sum_{\omega \in A} \text{pr}(\omega). \tag{1.2}$$

$\Pr(A)$ is called the *probability* of the event A, and the function Pr is called a *probability measure*. Note that we distinguish between "little pr" and "big Pr" which are related by (1.2) going one way and

$$\text{pr}(\omega) = \Pr(\{\omega\})$$

going the other way, where $\{\omega\}$ denotes the set containing just one point ω. Points and sets differ. Outcomes and events differ.

Theorem 1.2. *For any probability measure* Pr *on a sample space* Ω

$$0 \le \Pr(A) \le 1, \qquad A \subset \Omega.$$

Proof. That $\Pr(A) \ge 0$ for all A follows immediately from (1.2) and (1.1a). That $\text{pr}(A) \le 1$ for all A follows immediately from (1.1b) and (1.1a). □

So now we have two kinds of probabilities, probabilities of outcomes and probabilities of events, but for both we have by Theorems 1.1 and 1.2

Probabilities are between zero and one, inclusive.

This is a very important "sanity check." Whenever you calculate a probability, check that it is between zero and one. Whenever you derive a function, such as a PMF or a probability measure whose values are probabilities, check that all values of the function are between zero and one. This also goes for other functions that we will meet later (distribution functions). As we have already seen and will come to understand a lot more as we go along, probabilities of zero or one are quite special. So you should worry a bit when you calculate a probability that is exactly zero or exactly one. That is not obviously incorrect, but unless there is a special reason for the answer to come out that way, it is wrong too.

Example 1.2.1 (Discrete Uniform Distribution)**.**
Consider the discrete uniform distribution (Example 1.1.4) on the sample space $\{1, 2, 3, 4, 5, 6\}$.

In this course we emphasize probability mass functions over probability measures. We will often calculate probabilities of events $\Pr(A)$, but will rarely be interested in Pr as a function that maps sets A to real numbers $\Pr(A)$. So if the notion of a function whose argument ranges over sets rather than numbers has been bothering you, don't worry about it.

1.3 Random Variables and Expectation

A *random variable* is a real-valued function on the sample space (of a probability model). The *expectation* of a random variable X defined on a probability model with sample space Ω and PMF pr is

$$E(X) = \sum_{\omega \in \Omega} X(\omega) \operatorname{pr}(\omega).$$

From the definition of "random variable" it is clear that

> *any function of a random variable (or of random variables) is a random variable.*

If X is a random variable, then so is $g(X)$, by which we mean the function whose value at ω is $g[X(\omega)]$. Similarly, if X and Y are random variables, then so is $g(X, Y)$, by which we mean the function whose value at ω is $g[X(\omega), Y(\omega)]$.

We take time out for another pedantic digression on mathematical notation and terminology. We are careful about our notation for functions, which includes random variables. We never say $X(\omega)$ is a random variable, We say X is a random variable and $X(\omega)$ is its value at the outcome ω. Moreover, we never say x^2 is a function. If we want to indicate the squaring function, we have two options. We can define a function g by the formula $g(x) = x^2$, or if we don't need to introduce a symbol (g) for the function, we can write $x \mapsto x^2$, which is read "the function that maps x to x^2" or more tersely "x maps to x^2." Similarly, we never say e^x is a function. We say $x \mapsto e^x$ is a function, or if we are using

the notation $\exp(x)$ instead of e^x, we can say exp is a function and $\exp(x)$ is its value at the point x.

This carefulness (or pedantry if you prefer) is especially important when talking about the identity function $x \mapsto x$. In elementary mathematics, the identity function is hardly ever mentioned, but in order to make "a random variable is a function on the sample space" work as a definition, we will need the identity function.

If we have a probability model with sample space S that is a subset of the real numbers and and PMF f, and we consider the random variable X that is the identity function on the sample space (which is real-valued hence a random variable), then

$$E(X) = \sum_{x \in S} x f(x) \tag{1.3}$$

because $X(x) = x$ for all x. We also have for any real-valued function g on the sample space

$$E\{g(X)\} = \sum_{x \in S} g(x) f(x). \tag{1.4}$$

If you like, you can consider (1.3) and (1.4) as two separate definitions: (1.3) is for $E(X)$ where X is the "basic" random variable and (1.4) is for $E\{g(X)\}$, that is, for functions of the "basic" random variable. But a more sophisticated view is that (1.3) is the special case of (1.4) that arises when g is the identity function. So you don't need to memorize both (1.3) and (1.4). You only need to memorize (1.4) and understand that (1.3) is the special case you get when $g(x) = x$.

Better still, you should not memorize any equation, but understand the concept behind the equation, in the spirit of "mathematics is invariant under changes of notation." To calculate the expectation of a random variable you sum over the sample space, and each term in the sum is the product of the value of the random variable corresponding to an outcome and the probability of that outcome.

If we have a probability model with sample space $\{x_1, \ldots, x_n\}$ that is a subset of $\mathbb{R}$ and PMF g and if h is a function from $\mathbb{R}$ to $\mathbb{R}$, then we write

$$E\{h(X)\} = \sum_{i=1}^{n} h(x_i) g(x_i),$$

where, as before, X is the identity function on the sample space.

Example 1.3.1 (Discrete Uniform Distribution)**.**

Consider the discrete uniform distribution on the set $\{1, \ldots, n\}$.

$$\begin{aligned} E(X) &= \sum_{i=1}^{n} x f(x) \\ &= \sum_{i=1}^{n} i \cdot \frac{1}{n} \\ &= \frac{1}{n} \sum_{i=1}^{n} i \end{aligned}$$

Now we need to know what is the sum of the integers 1, ..., n.

There is a story about the famous mathematician Carl Friedrich Gauss. When he was still in elementary school, the teacher gave the class the problem of adding the numbers from 1 to 100, hoping to occupy them for a while, but young Gauss got the answer almost immediately. Presumably, he had figured out the following argument. Notice that $1 + 100 = 101$, and $2 + 99 = 101$, and so forth to $50 + 51 = 101$. There are 50 pairs of numbers that sum to 101, so the answer is $50 \times 101 = 5050$.

Generalizing this argument, to the case where n is any even positive integer, there are $n/2$ pairs that sum to $n + 1$ so the sum is $n(n + 1)/2$. The n odd case is a bit more work. If we try to pair up the integers there is one left over. Symmetry suggests that it should be the one in the middle. There are $(n-1)/2$ pairs that add to $n+1$, and the middle value $(n+1)/2$ is unpaired. Thus they sum to

$$\frac{n-1}{2} \cdot (n+1) + \frac{n+1}{2} = \frac{(n-1)(n+1)}{2} + \frac{n+1}{2} = \frac{n(n+1)}{2}$$

Surprisingly, the same formula works for both the odd and even case.

Returning to our calculation of the expecation,

$$E(X) = \frac{1}{n} \sum_{i=1}^{n} i = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}.$$

Theorem 1.3.

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}.$$

The proof is given in Example 1.3.1.

Appendix A

Greek Letters

Table A.1: Table of Greek Letters (Continued on following page.)

name	capital letter	small letter	pronunciation	sound
alpha	A	α	AL-fah	short a
beta	B	β	BAY-tah	b
gamma	Γ	γ	GAM-ah	g
delta	Δ	δ	DEL-tah	d
epsilon	E	ϵ	EP-si-lon	e
zeta	Z	ζ	ZAY-tah	z
eta	H	η	AY-tah	long a
theta	Θ	θ or ϑ	THAY-tah	soft th (as in thin)
iota	I	ι	EYE-oh-tah	i
kappa	K	κ	KAP-ah	k
lambda	Λ	λ	LAM-dah	l
mu	M	μ	MYOO	m
nu	N	ν	NOO	n
xi	Ξ	ξ	KSEE	x (as in box)
omicron	O	o	OH-mi-kron	o
pi	Π	π	PIE	p
rho	R	ρ	RHOH	rh[1]
sigma	Σ	σ	SIG-mah	s
tau	T	τ	TAOW	t
upsilon	Υ	υ	UP-si-lon	u

[1]The sound of the Greek letter ρ is not used in English. English words, like *rhetoric* and *rhinoceros* that are descended from Greek words beginning with ρ have English pronunciations beginning with an "r" sound rather than "rh" (though the spelling reminds us of the Greek origin).

Table A.2: Table of Greek Letters (Continued.)

name	capital letter	small letter	pronunciation	sound
phi	Φ	ϕ or φ	FIE	f
chi	X	χ	KIE	guttural ch[2]
psi	Ψ	ψ	PSY	ps (as in stops)[3]
omega	Ω	ω	oh-MEG-ah	o

[2]The sound of the Greek letter χ is not used in English. It is heard in the German *Buch* or Scottish *loch*. English words, like *chemistry* and *chorus* that are descended from Greek words beginning with χ have English pronunciations beginning with a "k" sound rather than "guttural ch" (though the spelling reminds us of the Greek origin).

[3]English words, like *pseudonym* and *psychology* that are descended from Greek words beginning with ψ have English pronunciations beginning with an "s" sound rather than "ps" (though the spelling reminds us of the Greek origin).

Appendix B

Sets and Functions

B.1 Sets

In mathematics, a *set* is a collection of objects thought of as one object. Any sort of object can be collected in a set. The objects in the set are called its *elements*. The notation $x \in S$ says that x is an element of the set S.

If A and B are sets, then A is a *subset* of B if every element of A is an element of B. The notation $A \subset B$ says that A is a subset of B.

Sets can be indicated by listing the elements in curly brackets $\{1, 2, 3, 4\}$. Sets can collect anything, not just numbers

$$\{1, 2, \pi, \text{cabbage}, \{0, 1, 2\}\}$$

One of the elements of this set is itself a set $\{0, 1, 2\}$. Most of the sets we deal with are sets of numbers or vectors, nothing this odd.

The *empty set* $\{\}$ is the only set that has no elements. Like the number zero, it simplifies a lot of mathematics, but isn't very interesting in itself. The empty set has its own special notation $\varnothing$.

Some very important sets also get their own special notation.

- $\mathbb{N}$ denotes the natural numbers $\{0, 1, 2, \ldots\}$.
- $\mathbb{Z}$ denotes the integers $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$.
- $\mathbb{R}$ denotes the real numbers.

Another notation for sets is the *set builder* notation

$$\{\, x \in S : \text{some condition on } x \,\}$$

denotes the set of elements of S that satisfy the specified condition. For example,

$$\{\, x \in \mathbb{R} : x > 0 \,\}$$

is the set of positive real numbers.

B.2 Intervals

Another important special kind of set is an *interval*. We use the notation

$$(a, b) = \{\, x \in \mathbb{R} : a < x < b \,\} \tag{B.1}$$
$$[a, b] = \{\, x \in \mathbb{R} : a \leq x \leq b \,\} \tag{B.2}$$
$$(a, b] = \{\, x \in \mathbb{R} : a < x \leq b \,\} \tag{B.3}$$
$$[a, b) = \{\, x \in \mathbb{R} : a \leq x < b \,\} \tag{B.4}$$

which assumes a and b are real numbers such that $a < b$.

We say an interval is *open* if it does not contain its endpoints, and we say it is *closed* if it does contain its endpoints. Thus (B.1) is *open*, (B.2) is *closed*, and (B.3) and (B.4) are neither open nor closed.

We also use the notation

$$(a, \infty) = \{\, x \in \mathbb{R} : a < x \,\} \tag{B.5}$$
$$[a, \infty) = \{\, x \in \mathbb{R} : a \leq x \,\} \tag{B.6}$$
$$(-\infty, b) = \{\, x \in \mathbb{R} : x < b \,\} \tag{B.7}$$
$$(-\infty, b] = \{\, x \in \mathbb{R} : x \leq b \,\} \tag{B.8}$$
$$(-\infty, \infty) = \mathbb{R} \tag{B.9}$$

which assumes a and b are real numbers.

For these intervals, $-\infty$ and $+\infty$ are not considered endpoints because they are not real numbers. Thus (B.5) and (B.7) are *open*, (B.6) and (B.8) are *closed*, and (B.9) is both open and closed.

B.3 Functions

A mathematical *function* is a rule that for each point in one set called the *domain* of the function gives a point in another set called the *codomain* of the function. Functions are also called *maps* or *mappings*.

Functions are often denoted by single letters, such as f, in which case the rule maps points x in the domain to values $f(x)$ in the codomain. Note the distinction: f is a function, $f(x)$ is the value of this function at the point x.

If X is the domain and Y the codomain of the function f, then to indicate this we write

$$f : X \to Y$$

or

$$X \xrightarrow{\ f\ } Y$$

To define a function, we may give a formula

$$f(x) = x^2, \qquad x \in \mathbb{R}.$$

Note that we indicate the domain in the formula.

The same function can be indicated more simply by $x \mapsto x^2$, read "x maps to x^2." But this does not indicate the domain, which must be indicated some other way.

If the domain is a small finite set, then we can define the function with a table

x	1	2	3	4
$f(x)$	1/10	2/10	3/10	4/10

Functions can map any set to any set

x	red	orange	yellow	green	blue
$f(x)$	tomato	orange	lemon	lime	corn

Many students aren't used to being careful about domains of functions, but they will have to start now. Consider the function f defined by

$$f(x) = \sqrt{x}, \qquad x \geq 0.$$

Without the domain indicated, the definition makes no sense. In this course, we will always insist that the definition of a function includes a description of the domain. If you don't indicate the domain, then you haven't finished the problem.

Functions can be indicated by notations other than letters. The exponential function

$$\mathbb{R} \xrightarrow{\exp} (0, \infty)$$

has values denoted $\exp(x)$ or e^x. The logarithmic function

$$(0, \infty) \xrightarrow{\log} \mathbb{R}$$

has values denoted $\log(x)$.

These functions are inverses of each other

$$\log\bigl(\exp(x)\bigr) = x, \qquad \text{for all } x \text{ in the domain of exp}$$
$$\exp\bigl(\log(x)\bigr) = x, \qquad \text{for all } x \text{ in the domain of log}$$

If you are used to distinguishing between base e and base 10 logarithms, calling one $\ln(x)$ and the other $\log(x)$, forget it. In this course, $\log(x)$ always means the base e logarithm, also called natural logarithm. Base 10 logarithms should never be used in probability and statistics; they only cause confusion when used.

Two kinds of functions that simplify a lot of mathematics, but aren't very interesting in themselves are constant functions and identity functions. For any constant c, the function $x \mapsto c$ can be defined on any set and is called a *constant* function. The function $x \mapsto x$ can be defined on any set and is called the *identity* function for that set.

We never say x^2 is a function. We must always write $x \mapsto x^2$ to indicate the squaring function. If you are in the habit of calling x^2 a function, then how can you describe identity and constant functions? Would you say x is a function? Would you say 2 is a function? Better to be pedantically correct and say $x \mapsto x^2$ so we can also say $x \mapsto x$ and $x \mapsto 2$.

Appendix C

Summary of Brand-Name Distributions

C.1 The Discrete Uniform Distribution

The Abbreviation DiscUnif(S).

Type Discrete.

Rationale Equally likely outcomes.

The Sample Space Any finite set S.

The Probability Mass Function

$$f(x) = \frac{1}{n}, \qquad x \in S,$$

where n is the number of elements in S.

Moments For the case in which the sample space is $S = \{1, \ldots, n\}$

$$E(X) = \frac{n+1}{2}$$
$$\text{var}(X) = \frac{n^2 - 1}{12}$$

C.2 The Bernoulli Distribution

The Abbreviation Ber(p)

Type Discrete.

Rationale Any zero-or-one-valued random variable.

The Sample Space The two-element set $\{0, 1\}$.

The Parameter Real number p such that $0 \le p \le 1$.

The Probability Mass Function

$$f(x) = \begin{cases} p, & x = 1 \\ 1 - p & x = 0 \end{cases}$$

Moments

$$E(X) = p$$
$$\operatorname{var}(X) = p(1 - p)$$

Addition Rule If X_1, ..., X_k are IID $\operatorname{Ber}(p)$ random variables, then $X_1 + \cdots + X_k$ is a $\operatorname{Bin}(k, p)$ random variable.

Relation to Other Distributions

$$\operatorname{Ber}(p) = \operatorname{Bin}(1, p)$$

Index