Lecture 5: Random variables and expectation

Relevant textbook passages:

Pitman [9]: Sections 3.1–3.2

Larsen–Marx [8]: Sections 3.3–3.5

Supplemental reading: **Ash [2]**: Sections 3.1–3.3

5.1 Random variables

Recall the following Definition 2.5.2.

5.1.1 Definition *A **random variable** on a probability space $(\Omega, \mathcal{F}, P)$ is a real-valued function on Ω, that is,*

$$X : \Omega \to \boldsymbol{R},$$

*which has the following **measurability** property: for every interval $I \subset \boldsymbol{R}$ the inverse image of I is an event. That is,*

$$\{\omega \in \Omega : X(\omega) \in I\} \text{ is an event.}$$

*A **random vector** $\boldsymbol{X} = (X_1, \ldots, X_n)$ is simply a finite-dimensional vector of random variables.*

Note that when the collection $\mathcal{F}$ of events consists of all subsets of Ω, then the measurability is automatically satisfied. It is difficult to construct examples of functions that do not have the mesurability property, so in what follows I may simply ignore the issue.

5.1.2 Remark An interpretation of random variables used by engineers is that they represent *measurements* on the state of a system. See, e.g., Robert Gray [6]. Another interpretation is that random variables are the source of **data**. That is, a random vector turns the outcome of a random experimental into a numerical datum that we can analyze using numerically.

5.2 The correspondence between indicator functions and events

Recall that the **indicator function** $\mathbf{1}_E$ of the event E is a random variable $\mathbf{1}_E : \Omega \to \boldsymbol{R}$ defined by

$$\mathbf{1}_E(\omega) = \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E. \end{cases}$$

There are several useful correspondences between operations on sets and operations on their indicator functions. The following proposition summarizes a few of them. The proof is easy, and is left as an exercise.

5.2.1 Proposition *We have the following identities. (Note that operations on indicator functions are performed pointwise.)*

Complements: $\mathbf{1}_{E^c} = 1 - \mathbf{1}_E$.

Unions: $\mathbf{1}_{E \cup F} = \max\{\mathbf{1}_E, \mathbf{1}_F\} = \mathbf{1}_E \vee \mathbf{1}_F$.

Intersections: $\mathbf{1}_{EF} = \min\{\mathbf{1}_E, \mathbf{1}_F\} = \mathbf{1}_E \wedge \mathbf{1}_F$. *Also,* $\mathbf{1}_{EF} = \mathbf{1}_E \cdot \mathbf{1}_F$.

Monotone Limits *For a sequence $E_1, \dots, E_n, \dots$, that is increasing, i.e., $E_n \subset E_{n+1}$, also written $E_n \nearrow$, we have*

$$\mathbf{1}_{\bigcup_n E_n} = \lim_{n\to\infty} \mathbf{1}_{E_n}.$$

For a sequence $E_1, \dots, E_n, \dots$, that is decreasing, i.e., $E_n \supset E_{n+1}$, also written $E_n \searrow$, we have

$$\mathbf{1}_{\cap_n E_n} = \lim_{n\to\infty} \mathbf{1}_{E_n}.$$

Sums: $\mathbf{1}_E + \mathbf{1}_F \geqslant \mathbf{1}_{E\cup F}$. *Events E and F are disjoint if and only if $\mathbf{1}_E + \mathbf{1}_F = \mathbf{1}_{E\cup F}$.*

Sums: $\sum_{i=1}^n \mathbf{1}_{E_i}(\omega)$ *is the count of the number of sets E_i to which ω belongs, that is,*

$$\sum_{i=1}^n \mathbf{1}_{E_i}(\omega) = \#\{i : \omega \in E_i\}.$$

5.3 The distribution of a random variable

A random variable X on the probability space $(\Omega, \mathcal{F}, P)$ induces a probability measure or distribution on the real line as follows. Given an interval I, we define

$$P_X(I) = P\left(\{\omega \in \Omega : X(\omega) \in I\}\right).$$

This gives us probabilities for intervals. We can extend this to probabilities of other sets, such as complements of intervals, countable unions of intervals, countable intersections of countable unions of intervals, etc. It turns out that the probabilities of the intervals pin down the probabilities on the entire **Borel σ-algebra**, denoted $\mathcal{B}$. (Recall Appendix 2.10⋆.) This result is known as the **Carathéodory Extension Theorem**, and may be found in many places, such as [1, Chapter 10]. This probability measure on the real line $\boldsymbol{R}$ is called the **distribution** of the random variable X.

Pitman [9]: § 3.1

5.3.1 Definition *The random variable $X\colon \Omega \to \boldsymbol{R}$ on the probability space $(\Omega, \mathcal{F}, P)$ creates a new probability space $(\boldsymbol{R}, \mathcal{B}, P_X)$, where $\mathcal{B}$ is the Borel σ-algebra, and the **distribution** of X, P_X, is defined for $B \in \mathcal{B}$ by P_X defined by*

$$P_X(B) = P(X \in B).$$

The virtue of knowing the distribution is that for many purposes we can ignore the original probability space and only worry about the distribution on the induced sample space $\boldsymbol{R}$. But be sure to read section 3.1 in Pitman [9], especially p. 146, on the difference between two variables being equal and having the same distribution:

A random variable is a function on a sample space, and a distribution is a probability measure on the real numbers. It is possible for two random variables to be defined on different sample spaces, but still have the same distribution. For example, let X be the indicator that is one if a coin comes up Tails, and Y be the indicator that a die is odd. Assuming both the coin and the die are "fair," X and Y will have the same distribution, namely each is equal to one with probability 1/2 and zero with probability 1/2, but they are clearly different random variables.

5.4 Discrete random variables

A random variable X is **simple** if the range of X is finite. A random variable X is **discrete** if the range of X is countable (finite or denumerably infinite).

For a discrete random variable, let x belong to the range of X. The **probability mass function** p_X is given by

$$p_X(x) = P(X = x)$$

It completely determines the distribution of X.

5.5 The cumulative distribution function of a random variable

Pitman [9]: § 4.5
Larsen–Marx [8]: p. 127, p. 137

5.5.1 Definition *The **cumulative distribution function** F_X of the random variable X defined on the probability space $(\Omega, \mathcal{F}, P)$ is the function $F_X \colon \boldsymbol{R} \to [0,1]$ defined by*

$$F_X(x) = P(X \leqslant x) = P_X(-\infty, x].$$

We often write

$$X \sim F$$

to mean that the random variable X has cumulative distribution function F.

N.B. Many authors whom I respect, for instance, C. Radikrishna Rao [10], Leo Breiman [3], and most of the French define the cumulative distribution function using the strict inequality $X < x$ rather than $X \leqslant x$.

5.5.2 Fact *The cumulative distribution function F_X is a nondecreasing, right continuous function, and satisfies $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.*

5.6 Examples

5.6.1 Bernoulli random variables

The **Bernoulli distribution** is a discrete distribution that generalizes coin tossing. A random variable X with a Bernoulli(p) distribution takes on two values: 1 ("success") and 0 ("failure").

The probability mass function is

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

Its probability mass function and cumulative distribution function are not very interesting.

5.6.2 Binomial random variables

The **Binomial(n, p) distribution** is the distribution of the number X of "successes" in n independent Bernoulli(p) trials.

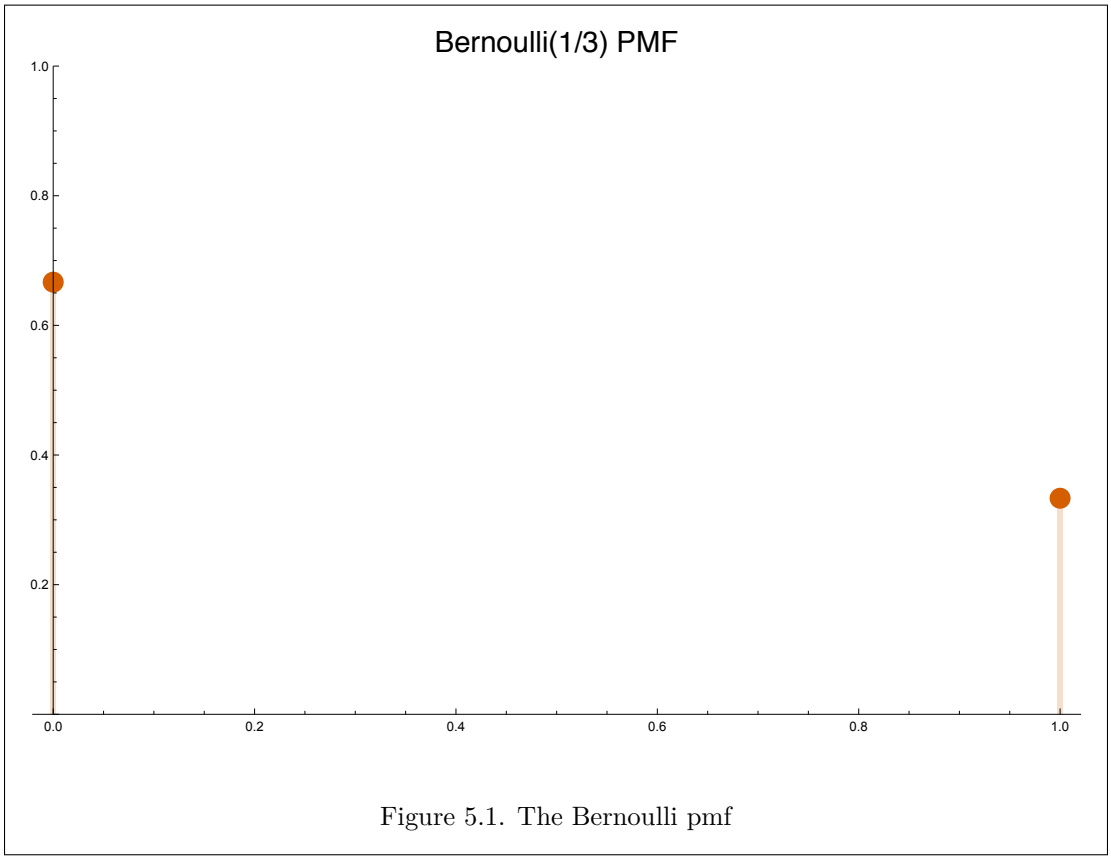

Figure 5.1. The Bernoulli pmf

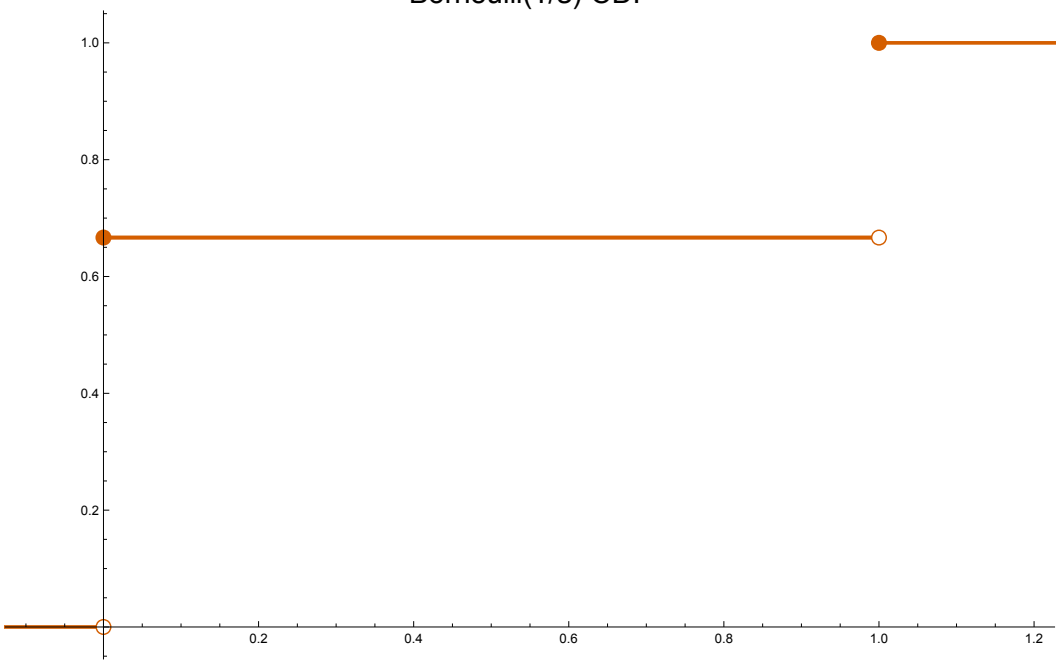

Figure 5.2. The Bernoulli cdf

The probability mass function is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, \ldots, n.$$

Note that the Binomial probability mass functions are **unimodal**. The **mode** is the value where the probability mass function assumes its maximum. Here this occurs at $X = pn$. When pn is not an integer, the mode(s) will be adjacent to pn. Note that the probability mass function for $p = 0.5$ is symmetric about pn, the height of the mode is lower, and the pmf is more "spread out." The probability mass functions for $p = 0.2$ and $p = 0.8$ are mirror images, which should be obvious from the formula for the probability mass function.

5.7 Parametrized distributions

Many, if not most, named probability distributions come in parametrized families. For instance, the binomial distribution has two parameters n and p; the Normal family of distributions (of which we shall hear a lot more later) has two parameters μ and σ (or σ^2).

The following definitions are standard, see, e.g, Forbes, et al. [5, p. 20].

5.7.1 Definition (Scale parameters) *Generally speaking, if a distribution $F = F(\cdot; \sigma)$ has a parameter $\sigma > 0$, where*

$$X \sim F(\cdot; 1) \quad \iff \quad \sigma X \sim F(\cdot; \sigma),$$

then σ is referred to a ***scale parameter****. However, on occasion some authors may refer to $1/\sigma$ as a scale parameter.*

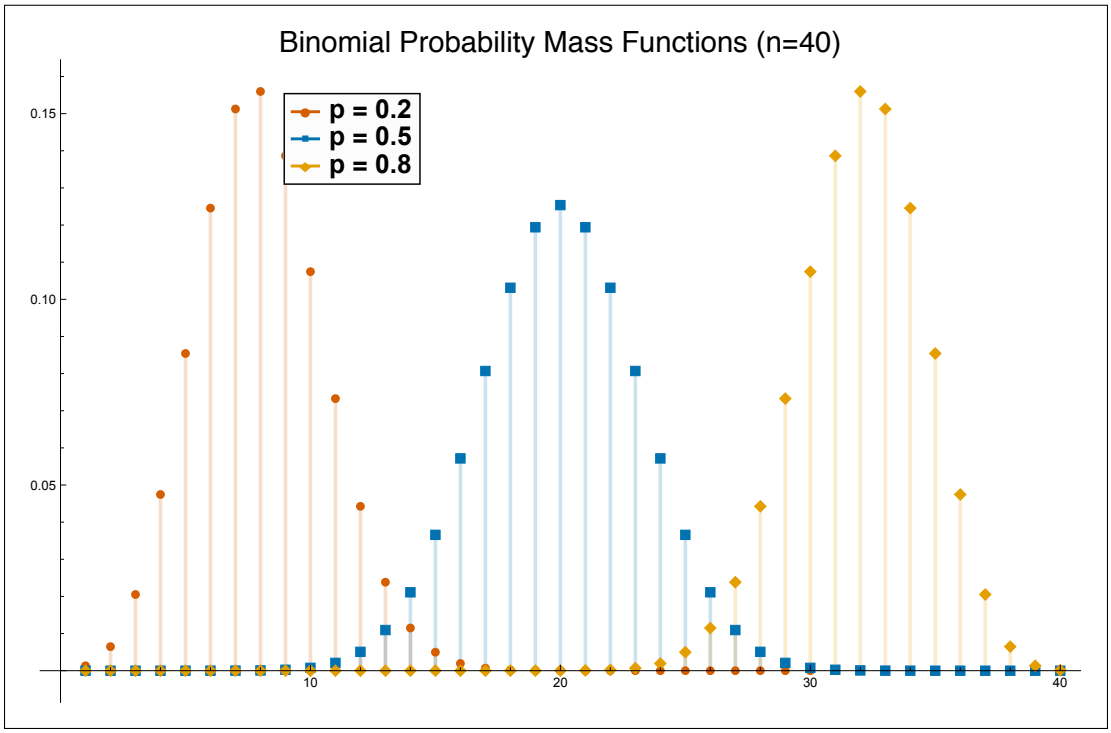

Figure 5.3. Binomial probability mass functions.

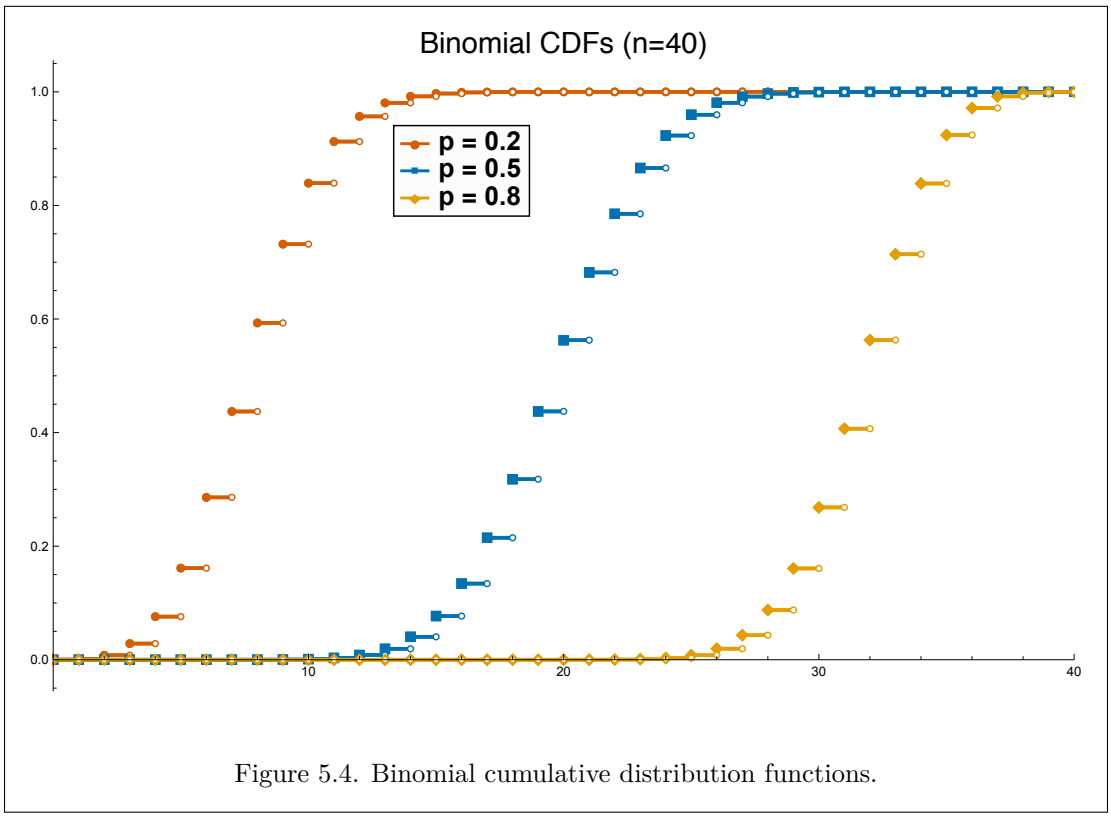

Figure 5.4. Binomial cumulative distribution functions.

5.7.2 Definition (Location parameters) *If a distribution $F = F(\cdot;\mu)$ has a parameter μ, such that if*

$$X \sim F(\cdot;0) \quad \iff \quad X + \mu \sim F(\cdot;\mu),$$

then μ is referred to a ***location parameter****.*

Other parameters may not have such simple interpretations. Sometimes they are called **shape parameters**. For the binomial distribution, the number of trials, n, is called the **size parameter**. Other parameters may be referred to as**degrees of freedom**.

5.8 ⋆ Stochastic dominance

Note: This material is in neither Pitman [9] nor Larsen–Marx [8].

Given two random variables X and Y, we say that X **stochastically dominates** Y if for every real number x

$$P(X \geqslant x) \geqslant P(Y \geqslant x),$$

and for some x this holds as a strict inequality. In other words, X stochastically dominates Y if for every x

$$F_X(x) \leqslant F_Y(x),$$

with a strict inequality for at least one x.

If X is the time to failure for one brand of hard drive, and Y is the time to failure for another, which hard drive do you want in your computer?

Note that the Binomial distributions for a fixed n are ordered so that a larger p stochastically dominates a smaller p. See Figure 5.4.

5.9 The expectation of a random variable on a finite state space

The expectation of a random variable is a concept that grew out of the study of gambling games. Suppose the sample space for a gambling game is the finite set

$$\Omega = \{\omega_1, \ldots, \omega_n\},$$

and that the probability of each outcome is given by the probability measure P on Ω. Suppose further that in outcome $\omega \in \Omega$, you win $X(\omega)$. What is a fair price to pay the casino to play this game? What the early probabilists settled on is what we now call the expectation of X.

5.9.1 Definition (Expectation on finite state spaces) *When X is a random variable on a probability space $(\Omega, \mathcal{F}, P)$ and Ω is finite, we define the* ***expectation*** $\boldsymbol{E}\,X$ *of X to be the number*

$$\sum_{\omega\in\Omega} X\omega P(\omega).$$

The case of a finite state space Ω covers most gambling situations, and most "classical" probability theory deals with this case. There are other terms for the expectation, including the **mean**, the **first moment**, or the {dfmathematical expectation, or even the **expectation operator**.

Why is this considered the "fair price?" For simplicity assume that each of n outcomes is equally likely (e.g., roulette). If we play the game n times and we get each possible out ω_i once, we shall have won $\sum_\omega X(\omega)$. So they argued the fair price per play should be $\sum_\omega X(\omega)/n = \boldsymbol{E}\,X$.

5.9.2 Example (Some simples examples of expectation) Here I'll use simple tables to describe random variables and their expectations.

Rademacher random variable

Toss a fair coin and let $Y = 1$ if Tails occurs and $Y = -1$ if Heads occurs. We can summarize this in the following table.

ω	$P(\omega)$	$X(\omega)$	$X(\omega)P(\omega)$
T	$1/2$	1	$1/2$
H	$1/2p$	-1	$-1/2$
		$\boldsymbol{E}\,X$	0

Bernoulli random variable There are two points in the sample space, Success and Failure. The Bernoulli random variable is 1 if a success occurs and 0 if a failure occurs. We can summarize this in the following table.

ω	$P(\omega)$	$X(\omega)$	$X(\omega)P(\omega)$
Success	p	1	p
Failure	$1-p$	0	0
		$\boldsymbol{E}\,X$	p

Binomial random variable This counts the number of success in n independent Bernoulli trials. There are 2^n points in the sample space. Here is a table for the case $n = 3$.

ω	$P(\omega)$	$X(\omega)$	$X(\omega)P(\omega)$
FFF	$(1-p)^3$	0	0
FFS	$(1-p)^2p$	1	$p(1-p)^2$
FSF	$(1-p)p(1-p)$	1	$p(1-p)^2$
FSS	$(1-p)p^2$	2	$2p^2(1-p)$
SFF	$p(1-p)^2$	1	$p(1-p)^2$
SFS	$p(1-p)p$	2	$2p^2(1-p)$
SSF	$p^2(1-p)$	2	$2p^2(1-p)$
SSS	p^3	3	$3p^3$
		$\boldsymbol{E}\,X$	$3p$

(The last step, adding up the last column to get $3p$ is tedious, but here goes: $p(1-p)^2 = p(1-2p+p^2) = p-2p^2+p^3$, and $p^2(1-p) = p^2-p^3$, so the sum of the last column becomes $3(p-2p^2+p^3)+6(p^2-p^3)+3p^3 = 3p$.)

Another way to do this would be to group all the points where $X = x$ to get

x	$(X = x)$	$P(X = x)$	$xP(X = x)$
0	$\{FFF\}$	$(1-p)^3$	0
1	$\{FFS, FSF, SFF\}$	$3(1-p)^2p$	$3\big(p-2p^2+p^3\big)$
2	$\{FSS, SFS, SSF\}$	$3(1-p)p^2$	$6\big(p^2-p^3\big)$
3	$\{SSS\}$	p^3	$3p^3$
		$\boldsymbol{E}\,X$	$3p$

This last calculation uses the probability mass function rather than the probability measure P on the state space.

Roll of a die Let X be the number showing on a standard die. Here is a table showing the

calculation of $\boldsymbol{E}\,X$.

ω	$P(\omega)$	$X(\omega)$	$X(\omega)P(\omega)$
1	1/6	1	1/6
2	1/6	2	2/6
3	1/6	3	3/6
4	1/6	4	4/6
5	1/6	5	5/6
6	1/6	6	6/6
		$\boldsymbol{E}\,X$	$\frac{21}{6} = 3\frac{1}{2}$

Roll of two dice Let X be the total showing on a roll of two standard dice. Here is a table showing the calculation of $\boldsymbol{E}\,X$, using the grouping method. Points in the state space are represented by a pair of digits, e.g., 62 means 6 on die 1 and 2 on die 2. Each point has probability 1/36

x	$(X = x)$	$p(x) = P(X = x)$	$xP(X = x)$
2	{11}	1/36	2/36
3	{12, 21}	2/36	6/36
4	{13, 22, 31}	3/36	12/36
5	{14, 23, 32, 41}	4/36	20/36
6	{15, 24, 33, 43, 51}	5/36	30/36
7	{16, 25, 34, 43, 52, 61}	6/36	42/36
8	{26, 35, 44, 53, 62}	5/36	40/36
9	{36, 45, 54, 63}	4/36	36/36
10	{46, 55, 64}	3/36	30/36
11	{56, 65}	2/36	22/36
12	{66}	1/36	12/36
		$\boldsymbol{E}\,X$	$\frac{252}{36} = 7$

□

We will see in Proposition 5.11.1 below that there are easier ways to calculate some of these expectations.

5.9.3 Remark Here is an interpretation of the expectation that you may find useful. At least it appears in many textbooks.

For a simple random variable X with values $x_1, \ldots, x_n$ imagine the real line as a massless balance beam with masses $p(x_i)$ placed at x_i for each i. Now place a fulcrum at the position μ. From what I recall of Ph 1a, the total torque on the beam is

$$\sum_i p(x_i)(x_i - \mu).$$

Which value of μ makes the total torque equal to zero? Since $\sum_i p(x_i) = 1$, it is easy to see that

$$\mu = \sum_i p(x_i)x_i$$

is the balancing point. That is, the beam is balanced at the expectation of X. In this sense, the expectation is the **location of the "center" of the distribution**.

Since the torque is also called the **moment** of the forces[1] the expectation is also known as the **first moment** of the random variable's distribution.

[1] According to my copy of the *OED* [11] the term "moment" comes from the Latin *momentum*, meaning "movement" or "moving force."

It follows that

$$\boldsymbol{E}\big(X - (\boldsymbol{E}\,X)\big) = 0.$$

Proof: By definition,

$$\boldsymbol{E}\,X = \sum_{\omega\in\Omega} X(\omega)P(\omega)$$

and

$$\begin{aligned}\boldsymbol{E}\big(X - (\boldsymbol{E}\,X)\big) &= \sum_{\omega\in\Omega}\big(X(\omega) - (\boldsymbol{E}\,X)\big)P(\omega)\\ &= \sum_{\omega\in\Omega} X(\omega)P(\omega) - \boldsymbol{E}\,X\underbrace{\sum_{\omega\in\Omega} P(\omega)}_{=1} = \boldsymbol{E}\,X - \boldsymbol{E}\,X = 0.\end{aligned}$$

∎

5.9.4 Remark We shall soon see that the expectation is the long run average value of X in independent experiments. This is known as the Law of Large Numbers, or more informally as the Law of Averages.

Interpretations of $\boldsymbol{E}\,X$:

- The "fair price" of a gamble X.
- The location of the "center of mass" of the distribution of X.
- Long run average value of X in independent experiments.
- If X is the indicator function of an event E, then $\boldsymbol{E}\,X$ is $P(E)$.

Here are two easily verified properties of expectation in this case.

1. If $X \geqslant 0$, then $\boldsymbol{E}\,X \geqslant 0$.
2. For real constants a, b,
$$\boldsymbol{E}(aX + bY) = a\,\boldsymbol{E}\,X + b\,\boldsymbol{E}\,Y.$$

The latter is just the distributive law:

$$\begin{aligned}E(aX + bY) &= \sum_{\omega\in\Omega}(aX(\omega) + bY(\omega))P(\omega)\\ &= a\sum_{\omega\in\Omega} X(\omega)P(\omega) + b\sum_{\omega\in\Omega} Y(\omega)P(\omega) = a\,\boldsymbol{E}\,X + b\,\boldsymbol{E}\,Y.\end{aligned}$$

Together these two properties are summarized by saying that **expectation is a positive linear operator**. We want to extend the definition to more general random variables in a way that preserves these properties.

5.10 Expectation of a simple random variable

Pitman [9]: § 3.1
Larsen–Marx [8]: § 3.5

We define the expectation in steps, starting with the expectation of a simple random variable.

5.10.1 Definition *Let X be a simple random variable on the probability space $(\Omega, \mathcal{F}, P)$ that takes on the distinct values $\{x_1, \dots, x_n\}$. The* ***canonical representation*** *or* ***standard representation*** *of X is given by*

$$X(\omega) = \sum_{i=1}^{n} x_i \mathbf{1}_{A_i}(\omega),$$

where $A_i = \{\omega \in \Omega : X(\omega) = x_i\} = (X = x_i)$.

Pitman [9]: § 3.2

5.10.2 Definition (Expectation of a simple random variable) *Let X be a simple random variable on the probability space $(\Omega, \mathcal{F}, P)$ with canonical representation*

$$X(\omega) = \sum_{i=1}^{n} x_i \mathbf{1}_{A_i}(\omega).$$

The ***expectation*** *of X is defined to be the number*

$$\boldsymbol{E}\, X = \sum_{i=1}^{n} x_i P(A_i).$$

Consequently, the expectation $\boldsymbol{E}\, \mathbf{1}_E$ of the indicator function $\mathbf{1}_E$ of an event E is $P(E)$.

Note that the expectation of a simple random variable X is determined by its distribution on $\boldsymbol{R}$.

5.10.3 Proposition *For a simple random variable X,*

$$\boldsymbol{E}\, X = \sum_{x \in \operatorname{range} X} x p_X(x),$$

where p_X is the probability mass function for X.

If X is a simple random variable on a discrete probability space $(\Omega, \mathcal{F}, P)$, then

$$\boldsymbol{E}\, X = \sum_{\omega \in \Omega} X(\omega) P(\omega).$$

Proof: The first expression is just a rewriting of the canonical representation, and the second follows from the canonical representation by replacing $P(A_i)$ with $\sum_{\omega \in A_i} P(\omega)$. ∎

In other words the expectation is a weighted average of the values of X where the weights are the probabilities attached to those values.

5.10.4 Remark If you are alert and looking for trouble, you might point out that a simple random variable X can be written in many ways as a linear combination of indicator functions.

The canonical way is to enumerate the range of X as $\{x_1, \dots, x_n\}$, and set $E_i = \{\omega : X(\omega) = x_i\}$, $i = 1, \dots, n$. Then

$$X = \sum_{i=1}^{n} x_i \mathbf{1}_{E_i}.$$

But there is usually more than one way to skin a cat. For instance, let $S = \{0, 1\}$ and consider the two linear combinations of indicator functions:

$$X = \mathbf{1}_S - \mathbf{1}_{\{0\}} = \mathbf{1}_{\{1\}}.$$

How do we know that the so-called definition of expectation will give the same result for each representation? And if it doesn't, what makes the canonical representation so special, anyhow?

The answer is given in the next proposition.

5.10.5 Proposition (Expectation does not depend on the representation) *Let X be a simple random variable with representations*

$$X = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i} = \sum_{j=1}^{m} b_j \mathbf{1}_{B_j}.$$

Then

$$\boldsymbol{E}\, X = \sum_{i=1}^{n} a_i P(A_i) = \sum_{j=1}^{m} b_j P(B_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} P(A_i B_j),$$

where $c_{ij} = a_i = b_j$ if $A_i B_j \neq \varnothing$, and $c_{ij} = 0$ otherwise.

The proof of the proposition is an exercise in keeping track of your notation and using the additivity property of probability, and I leave it as an exercise for the masochistic.

5.10.6 Remark Note that $\boldsymbol{E}$ can be (and is) regarded as an **operator** on the space of simple random variables. That is, it assigns to each random variable X a real number $\boldsymbol{E}\, X$. It is customary to write operators without parentheses, that is, as $\boldsymbol{E}\, X$ instead of $\boldsymbol{E}(X)$ (although Pitman uses parentheses). This practice can be a little ambiguous. For instance, if X is a random variable, so is X^2, so what does $\boldsymbol{E}\, X^2$ mean? Is it $\boldsymbol{E}(X^2)$ or $(\boldsymbol{E}\, X)^2$? The answer is $\boldsymbol{E}(X^2)$, the operator applied to the random variable X^2. Similarly, most people write $\boldsymbol{E}\, XY$ instead of $\boldsymbol{E}(XY)$, etc. (In a programming class you would say the operator $\boldsymbol{E}$ has lower precedence than arithmetic operators.) There are a few expressions coming up where I may add extra parentheses for clarity.

5.11 Expectation of a sum of simple random variables

5.11.1 Proposition *The expectation of the sum of two random variables is the sum of their expectations.*

Proof: If X and Y are simple random variables on the state space Ω with range $X = \{x_1, \dots, x_m\}$ and range $Y = \{y_1, \dots, y_n\}$, then $X + Y$ is a simple random variable with range$(X + Y) \subset \{x_i + y_j : i = 1, \dots, m,\ j = 1, \dots, n\}$. Now let $E_i = (X = x_i)$ and $F_j = (Y = y_j)$. Then the E_i's partition Ω and the F_j's partition Ω. So for each $j = 1, \dots, n$ we have

$$F_j = \bigcup_{i=1}^{m} E_i F_j, \quad \text{and so} \quad P(F_j) = \sum_{i=1}^{m} P(E_i F_j)$$

and for each $i = 1, \dots, m$ we have

$$E_i = \bigcup_{j=1}^{n} E_i F_j, \quad \text{and so} \quad P(E_i) = \sum_{j=1}^{n} P(E_i F_j)$$

By Definition 5.10.2 and Proposition 5.10.5,

$$\boldsymbol{E}(X+Y) = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_i+y_j)P(X=x_i \;\&\; Y=y_j) = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_i+y_j)P(E_iF_j).$$

Rearranging the right-hand side using the distributive law gives

$$\begin{aligned}\boldsymbol{E}(X+Y) &= \sum_{i=1}^{m}\sum_{j=1}^{n}x_iP(E_iF_j) + \sum_{i=1}^{m}\sum_{j=1}^{n}y_jP(E_iF_j)\\ &= \sum_{i=1}^{m}x_i\sum_{j=1}^{n}P(E_iF_j) + \sum_{j=1}^{n}y_j\sum_{i=1}^{m}P(E_iF_j)\\ &= \sum_{i=1}^{m}x_iP(E_i) + \sum_{j=1}^{n}y_jP(F_j) = \boldsymbol{E}\,X + \boldsymbol{E}\,Y.\end{aligned}$$

∎

This proposition can be used to simplify some of the expectations in Example 5.9.2.

5.12 Expectation of a function of a simple random variable

If X with canonical representation

$$X = \sum_{i=1}^{n} x_i\mathbf{1}_{A_i}$$

is a simple random variable on a probability space $(\Omega, \mathcal{F}, P)$ and g is a function from $\boldsymbol{R}$ to $\boldsymbol{R}$, then the composition $g \circ X$ is also a simple random variable,

$$g \circ X = \sum_{i=1}^{n} g(x_i)\mathbf{1}_{A_i}.$$

Then

$$\begin{aligned}\boldsymbol{E}(g\circ X) &= \sum_{i=1}^{n} g(x_i)P(A_i),\\ &= \sum_{x\in\operatorname{range} X} g(x)p_X(x)\end{aligned}$$

5.12.1 Extended random variables

We shall presently see that we may want to allow random variables to assume the extended-real values ∞ or $-\infty$.

5.12.1 Definition (Expectation for simple random variables) *For extended-real valued simple random variables, we adopt the convention that $\infty \cdot 0 = (-\infty) \cdot 0 = 0$.*

- *If $P(X = \infty) = P(X = -\infty) = 0$, then $\boldsymbol{E}\, X$ is computed as above, and is finite.*
- *If $P(X = \infty) > 0$ and $P(X = -\infty) = 0$, then $\boldsymbol{E}\, X$ is ∞.*
- *If $P(X = \infty) = 0$ and $P(X = -\infty) > 0$, then $\boldsymbol{E}\, X$ is $-\infty$.*
- *If $P(X = \infty) > 0$ and $P(X = -\infty) > 0$, then $\boldsymbol{E}\, X$ **does not exist**.*

5.13 The expectation of a nonnegative random variable

We now want to define the expectation for a random variable that is not simple. For instance, a geometric random variable, which is the number of independent Bernoulli trials it takes to get to the first success is not simple, but it is discrete. Proposition 5.10.3 suggests that for a discrete random variable X we should define $\boldsymbol{E}$ to be $\sum_{x \in \operatorname{range} X} x p_X(x)$. We will almost do this, but we take a slightly sideways approach. We move on to, not general discrete random variables, but to (not necessarily discrete) random variables taking on values in the range $[0, \infty]$. (Note that simple random variables, by definition, take on only values in $\boldsymbol{R}$, and cannot take on the value ∞.)

5.13.1 Definition (Expectation for nonnegative random variables) *Let X be a random variable on the probability space $(\Omega, \mathcal{F}, P)$ and taking on values in $[0, \infty]$. Then define*

$$\boldsymbol{E}\, X = \sup\{\boldsymbol{E}\, Y : Y \text{ is a simple random variable and } Y \leqslant X \text{ a.s.}\}.$$

(Recall that E a.s. means $P(E) = 1$.)

Note that this allows for $\boldsymbol{E}\, X$ to be the value ∞. We have the following result that can be used to find $\boldsymbol{E}\, X$.

5.13.2 Monotone Convergence Theorem *Let X be a nonnegative extended real-valued random variable on the probability space $(\Omega, \mathcal{F}, P)$. Let $X_n \geqslant 0$ be an increasing sequence of nonnegative random variables on Ω, that is,*

$$0 \leqslant X_1 \leqslant X_2 \leqslant \cdots \leqslant X_n \cdots,$$

(where the inequality holds pointwise for each $\omega \in \Omega$) and assume

$$X = \lim_n X_n = \sup_n X_n.$$

Then

$$\boldsymbol{E}\, X = \lim_n \boldsymbol{E}\, X_n.$$

(The possibility $\boldsymbol{E}\, X = \infty$ is allowed.)

This is a standard result and may be found, for instance, in Breiman [3, A.26, p. 397] or Halmos [7, Theorem B, p. 112].

5.13.3 Remark Given a nonnegative random variable X, a standard way of constructing an increasing sequence $X_1, X_2, \dots$ of simple random variables such that $X = \lim_n X_n = \sup_n X_n$ is this. For each n, define X_n by

$$X_n(\omega) = \begin{cases} \frac{k}{2^n} & \text{if } \frac{k}{2^n} \leqslant X(\omega) < \frac{k+1}{2^n},\ k = 0, \dots, 2^{2n} \\ 2^n & X(\omega) \geqslant 2^n. \end{cases}$$

This works even if $X(\omega) = \infty$.

This procedure amount to taking the interval $[0, 2^n)$ and dividing it into 2^{2n} intervals $I_k = [(k-1)/2^n, k/2^n)$ of length $1/2^n$, for $k = 1, \dots, 2^{2n}$. Then for each point $s \in X^{-1}[I_k]$, the inverse image of I_k, we set $X_n(\omega) = (k-1)/2^n$, the bottom end of the interval I_k. For $s \in X^{-1}\big[[2^n, \infty)\big]$ we set $X_n(\omega) = 2^n$. See Figure 5.5.

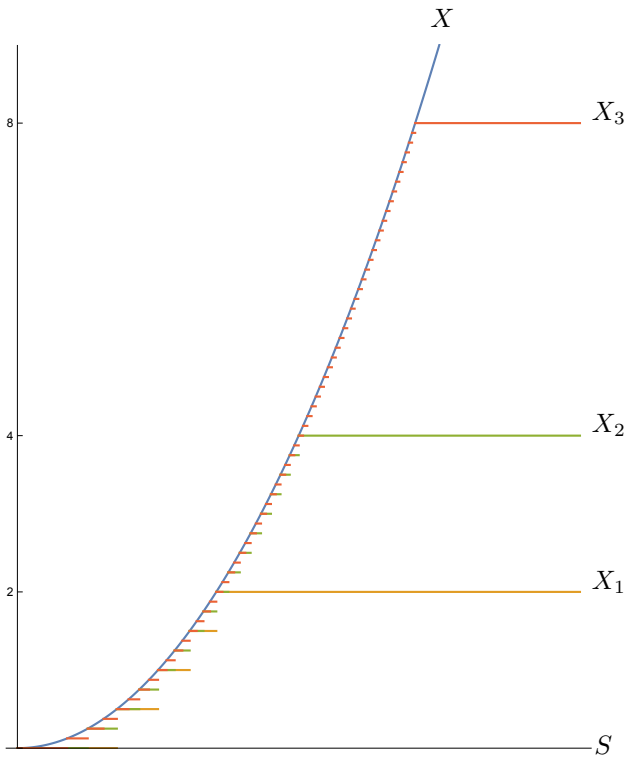

Figure 5.5. Approximating a nonnegative function X by simple functions X_1, X_2, X_3, etc., as described in Remark 5.13.3.

A corollary of the Monotone Convergence Theorem offers a simple way to calculate the expectation of discrete nonnegative random variable.

5.13.4 Corollary (Expectation as an infinite series) *Let X be a discrete nonnegative random variable with range $\{x_1, x_2, \dots\}$ and probability mass function p. Then*

$$\boldsymbol{E}\, X = \sum_{i=1}^{\infty} x_i p(x_i).$$

Proof: Let X_n be the simple random variable $X_n = \sum_{i=1}^{n} x_i \mathbf{1}_{(X=x_i)}$. (Note that $X_n(\omega) = 0$ unless $X(\omega) \in \{x_1, \dots, x_n\}$.) Then X_n is an increasing sequence of random variables with

$X_n \nearrow X$ and

$$\boldsymbol{E}\, X_n = \sum_{i=1}^{n} x_i p(x_i),$$

so by the Monotone Convergence Theorem 5.13.2,

$$\boldsymbol{E}\, X = \lim_{n\to\infty} \sum_{i=1}^{n} x_i p(x_i) = \sum_{i=1}^{\infty} x_i p(x_i).$$

∎

5.14 The St. Petersburg Paradox

There is at least one problem with the interpretation of expectation as a fair price.

5.14.1 Example (The St. Petersburg Paradox) Consider the following game: Toss a fair coin until the first Heads appears. If this happens on n^{th} toss, you win 2^n.

What is the expected value of this game?

The sample space S, for this experiment was discussed in Example 1.4.4,

$$S = \{H, \quad TH, \quad TTH, \quad \ldots, \quad \underbrace{TT\cdots T}_{n-1}H, \quad \ldots, \quad \overline{TTTT\cdots}\,\},$$

and

$$X(\underbrace{TT\cdots T}_{n-1}H) = 2^n, \qquad \text{and} \qquad X(\overline{TTTT\cdots}) = 0.$$

Now $P(\underbrace{TT\cdots T}_{n-1}H) = 1/2^n$, so by Corollary 5.13.4 of the Monotone Convergence Theorem 5.13.2,

$$\boldsymbol{E}\, X = \sum_{n=1}^{\infty} 2^n \frac{1}{2^n} = \sum_{n=1}^{\infty} 1 = \infty.$$

So if the expectation is a fair price, you should be willing to pay *any* price to play this game.

But wait! What is the probability that the game stops in a finite number of tosses? Let E_n be the event that the first Tails occurs on toss n. The event that the game stops in finitely many tosses is the countable disjoint union $\bigcup_{n=1}^{\infty} E_n$. (Do you see why?) But this has probability $\sum_{n=1}^{\infty} 1/2^n = 1$. So with probability 1 the game will end for some n, and you will receive $2^n < \infty$. This was regarded at the time as a paradox.

We shall see later that the reason expectation is not a good measure of "fairness" in this case is that the "Law of Averages" breaks down for random variables that do not have a finite expectation. □

Aside: According to Diaconis and Skyrms [4], the paradox was first posed by Nicholas Bernoulli in a letter to Pierre Raymond de Montmort on September 9, 1713. it was "resolved" independently by Gabriel Cramer and Nicholas's brother Daniel Bernoulli. Daniel, a one-time resident of the eponymous Russian city, published his arguments in the *Commentaries of the Imperial Academy of Science of Saint Petersburg* (1738).

5.14.2 Remark The expected length of a St. Petersburg game is

$$\sum_{n=1}^{\infty} (\text{length of game if first Tails is on toss } n) \times \text{Prob}\,(\text{first Tails is on toss } n) = \sum_{n=1}^{\infty} n2^{-n} = 2.$$

For a derivation of the value of this series, see Proposition S1.2.1 in Supplement 1.

5.15 The expectation of a general random variable

So far we have defined the expectation of a nonnegative random variable. How do we extend this to more general random variables? The answer is to split the random variable into two parts. Given a random variable X, define the random variables X^+ and X^- by

$$X^+ = \max\{X, 0\} \quad \text{and} \quad X^- = \max\{-X, 0\}.$$

Then X^+ agrees with X when $X \geqslant 0$, and $X^+ = 0$ when $X \leqslant 0$. Similarly, $X^- = -X$ when $X \leqslant 0$, and $X^- = 0$ when $X \geqslant 0$. We call X^+ the **positive part** of X, and X^- the **negative** part of X. N.B. The negative part X^- of X is actually a nonnegative random variable. We have the following identities:

$$X = X^+ - X^- \quad \text{and} \quad |X| = X^+ + X^-.$$

Since X^+ and X^- are nonnegative random variables, we know how to take their expectations. We now define the expectation of X in terms of the expectation of its positive and negative parts.

5.15.1 Definition (Expectation of a general random variable) *Let X be a random variable. Then we define*

$$\boldsymbol{E}\, X = \boldsymbol{E}\, X^+ - \boldsymbol{E}\, X^-,$$

except that when

$$\boldsymbol{E}\, X^+ = \boldsymbol{E}\, X^- = \infty, \text{ we say that } \boldsymbol{E}\, X \textbf{ does not exist}.$$

So

- *If $\boldsymbol{E}\, X^+ = \infty$ and EX^- is finite, then $\boldsymbol{E}\, X = \infty$, and we say that X has **infinite expectation**.*
- *If $\boldsymbol{E}\, X^- = \infty$ and EX^+ is finite, then $\boldsymbol{E}\, X = -\infty$, and we say that X has **negatively infinite expectation**.*

It follows that

5.15.2 Proposition *The expectation $\boldsymbol{E}\, X$ is finite if and only if $\boldsymbol{E}\, |X|$ is finite.*

We have just seen that if the sample space is infinite, it is possible to construct random variables whose expectation is a divergent series, that is, the expectation is infinite.

In terms of our balance beam interpretation of expectation, if we put a mass of 2^n at the position $1/2^n$ on the beam, for each $n = 1, 2, \ldots$, then there is no finite mass that we can put anywhere, no matter how far to the left, to get the beam to balance. You might say that's because we have an infinite mass on the right-hand side of the beam, but it's more subtle. Suppose I put only a mass of one at each position $1/2^n$. Then a single unit of mass at position -1 would balance the beam.

You might wonder if any "naturally occurring" random variables have infinite expectation, or if they only exist in the demented minds of mathematicians. The answer, unfortunately, is yes. Take a random walk that starts at zero, and at each clock tick a step of size ± 1 is taken with equal probability. We shall see in Supplement 7 that the number of periods we have to wait

to return to zero is a random variable with infinite expectation. During the 2017 Rose Bowl game, I was talking with a colleague in econometrics about a nonparametric estimation problem for latent variables in which some his terms were random variables with infinite expectations. So yes, there are random variables that pop up in practice, and have infinite expectation.

There are worse problems that can result. Imagine the following variant of the St. Petersburg Paradox. First roll a fair die. If it comes up even, then play the standard St. Petersburg game: If the first Tails happens on n^{th} toss, you win 2^n. if the die comes up odd, then if the first Tails happens on n^{th} toss, you *lose* 2^n. Thus you win 2^n with probability 2^{n+1} and "win" -2^n with probability 2^{n+1}, so the expectation is the infinite series

$$\sum_{n=1}^{\infty}(2^n - 2^n)/2^{n+1} = \frac{1}{2} - \frac{1}{2} + \frac{1}{2} - \frac{1}{2} + \cdots,$$

which is not an absolutely convergent series, so the expectation of the random variable **does not exist**.

You might say that the expectation of the random variable above should be defined to be zero. But when we get to the Law of Large Numbers (the law of averages) in Lecture 7, we shall see that this is not a useful notion of expectation.

5.16 Independent random variables

Pitman [9]: pp. 151–154

5.16.1 Definition *The pair X, Y of random variables is **stochastically independent random variables** if for every $B_1, B_2 \subset \boldsymbol{R}$,* [a]

$$P(X \in B_1 \text{ and } Y \in B_2) = P(X \in B_1) \cdot P(Y \in B_2).$$

*More generally, a set $\mathcal{X}$ of random variables is **mutually stochastically independent** if for every finite subset of random variables $X_1, \ldots, X_n$ of $\mathcal{X}$ and every collection $B_1, \ldots, B_n$ of subsets*[1] *of $\boldsymbol{R}$,*

$$P(X_1 \in B_1, \ldots, X_n \in B_n) = P(X_1 \in B_1) \cdots P(X_n \in B_n).$$

[a] Caveat: B_i must be a Borel set.

5.16.2 Example (Pairwise independence does not imply independence) Let X and Y be independent Bernoulli(1/2) random variables (coin tosses), and let Z be the parity of $X + Y$. (That is, $Z = 1$ if $X + Y$ is odd, and 0 otherwise) Then X, Y, and Z are pairwise stochastically independent (any pair are independent); but X, Y, Z are *not* mutually stochastically independent.

You will be asked to prove this in the homework. □

5.16.3 Definition *A sequence $X_1, X_2, \ldots$ (finite or infinite) is **independent and identically distributed**, abbreviated **i.i.d.**, if they have a common distribution function and are stochastically independent.*

5.16.1 Functions of independent random variables

5.16.4 Theorem *If $X_1, \dots, X_n$ are independent random variables and $f_1, \dots, f_n$ are (Borel) functions, then*

$$f_1(X_1), \dots, f_n(X_n) \text{ are independent random variables.}$$

Proof: We need to show that every collection $B_1, \dots, B_n$ of (Borel) subsets of $\boldsymbol{R}$,

$$P(f_1(X_1) \in B_1, \dots, f_n(X_n) \in B_n) = P(f(X_1) \in B_1) \cdots P(f_n(X_n) \in B_n),$$

but

$$\begin{aligned} P(f_1(X_1) \in B_1, \dots, f_n(X_n) \in B_n) &= P\big(X_1 \in f_1^{-1}[B_1], \dots, X_n \in f_n^{-1}[B_n]\big) \\ &= P\big(X_1 \in f_1^{-1}[B_1]\big) \cdots P\big(X_n \in f_n^{-1}[B_n]\big) \\ &= P(f(X_1) \in B_1) \cdots P(f_n(X_n) \in B_n). \end{aligned}$$

∎

Bibliography

[1] C. D. Aliprantis and K. C. Border. 2006. *Infinite dimensional analysis: A hitchhiker's guide*, 3d. ed. Berlin: Springer–Verlag.

[2] R. B. Ash. 2008. *Basic probability theory*. Mineola, New York: Dover. Reprint of the 1970 edition published by John Wiley and Sons.

[3] L. Breiman. 1968. *Probability*. Reading, Massachusetts: Addison Wesley.

[4] P. Diaconis and B. Skyrms. 2018. *Ten great ideas about chance*. Princeton, New Jersey: Princeton University Press.

[5] C. Forbes, M. Evans, N. Hastings, and B. Peacock. 2011. *Statistical distributions*, 4th. ed. Hoboken, New Jersey: John Wiley & Sons.

[6] R. M. Gray. 1988. *Probability, random processes, and ergodic properties*. New York: Springer–Verlag.

[7] P. R. Halmos. 1974. *Measure theory*. Graduate Texts in Mathematics. New York: Springer–Verlag. Reprint of the edition published by Van Nostrand, 1950.

[8] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[9] J. Pitman. 1993. *Probability*. Springer Texts in Statistics. New York, Berlin, and Heidelberg: Springer.

[10] C. R. Rao. 1973. *Linear statistical inference and its applications*, 2d. ed. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

[11] J. A. Simpson and E. S. C. Weiner, eds. 1989. *The Oxford English Dictionary*, 2d. ed. Oxford: Oxford University Press.