Caltech Department of Mathematics

Ma 3/103 KC Border
Introduction to Probability and Statistics Winter 2017

# Lecture 5: Random variables and expectation

**Relevant textbook passages:**
**Pitman [5]:** Sections 3.1–3.2
**Larsen–Marx [4]:** Sections 3.3–3.5

## 5.1 Random variables

**5.1.1 Definition** *A **random variable** on a probability space $(S, \mathcal{E}, P)$ is a real-valued function on $S$ which has the property that for every interval $I \subset \boldsymbol{R}$ the inverse image of $I$ is an event.*

Note that when the collection $\mathcal{E}$ of events consists of all subsets of $S$, then the requirement that inverse images of intervals be events is automatically satisfied.

**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 [3].

There is another definition of random variable that is quite common, especially in electrical engineering.

**5.1.3 Definition (Another kind of random variable)** *Given a set $\mathcal{A}$ of* symbols *or* letters, *called the **alphabet**, a random variable is defined to be a function from $S$ into $\mathcal{A}$.*

While we could enumerate the symbols in the alphabet and treat the random variable as a real-valued function, the arithmetic operations have no significance: what letter is the sum of the letters A and B?

Traditionally, probabilists and statisticians use upper-case Latin letters near the end of the alphabet to denote random variables. This has confused generations of students, who have trouble thinking of random variables as functions. For the sake of tradition, and so that you get used to it, we follow suit. So a **random variable** $X$ is a function

$$X\colon S \to \boldsymbol{R} \quad \text{such that for each interval } I, \quad \{s \in S : X(s) \in I\} \in \mathcal{E}.$$

We shall adopt the following notational convention, which I refer to as **statistician's notation**, that

$$(X \in I) \text{ means } \{s \in S : X(s) \in I\}.$$

Likewise $(X \leqslant t)$ means $\{s \in S : X(s) \leqslant t\}$, etc.

If $E$ belongs to $\mathcal{E}$, then its **indicator function** $\mathbf{1}_E$, defined by

$$\mathbf{1}_E(s) = \begin{cases} 0 & s \notin E \\ 1 & s \in E, \end{cases}$$

is a random variable.

## 5.2 The correspondence between indicator functions and events

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** *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*

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

*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*

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

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

*Also note that $\sum_{i=1}^n \mathbf{1}_{E_i}(s)$ is the count of the number of sets $E_i$ to which $s$ belongs, i.e., $\sum_{i=1}^n \mathbf{1}_{E_i}(s) = |\{i : s \in E_i\}|$.*

## 5.3 The distribution of a random variable

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

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

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.[1] This probability measure on the real line $\boldsymbol{R}$ is called the **distribution** of the random variable $X$.

**Pitman [5]:** § 3.1

**5.3.1 Definition** *The **distribution** of the random variable $X\colon S \to \boldsymbol{R}$ on the probability space $(S, \mathcal{E}, P)$ is the probability measure $P_X$ defined on $\boldsymbol{R}$ by*

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

The virtue of knowing the distribution is that for many purposes we can ignore the probability space and only worry about the distribution. But be sure to read section 3.1 in Pitman [5], especially p. 146, on the difference between two variables being equal and having the same distribution:

[1] It turns out that the probabilities of the intervals pin down the probabilities on a whole $\sigma$-algebra of subsets of real numbers, called the **Borel $\sigma$-algebra**. This result is known as the **Carathéodory Extension Theorem**, and may be found in many places, such as [1, Chapter 10]. Sets that belong to the Borel $\sigma$-algebra are called **Borel sets**. Every interval, every open set, and every closed set belongs to this $\sigma$-algebra. In fact, you need to take an advanced analysis class to be able to describe a set that is not a Borel set. (This is beginning to sound like a broken record. Oops! Have you ever even heard a broken record?)

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).

## 5.5 The probability mass function

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 $X$.

## 5.6 The cumulative distribution function

**Pitman [5]:** § 4.5
**Larsen–Marx [4]:** p. 127, p. 137

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

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

**N.B.** Many authors whom I respect, for instance, C. Radikrishna Rao [6], Leo Breiman [2], and most of the French define the cumulative distribution function using the strict inequality $<$ rather than $\leqslant$.

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

We often write

$$X \sim F$$

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

## 5.7 Examples

### 5.7.1 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 pmf and cdf are not very interesting.

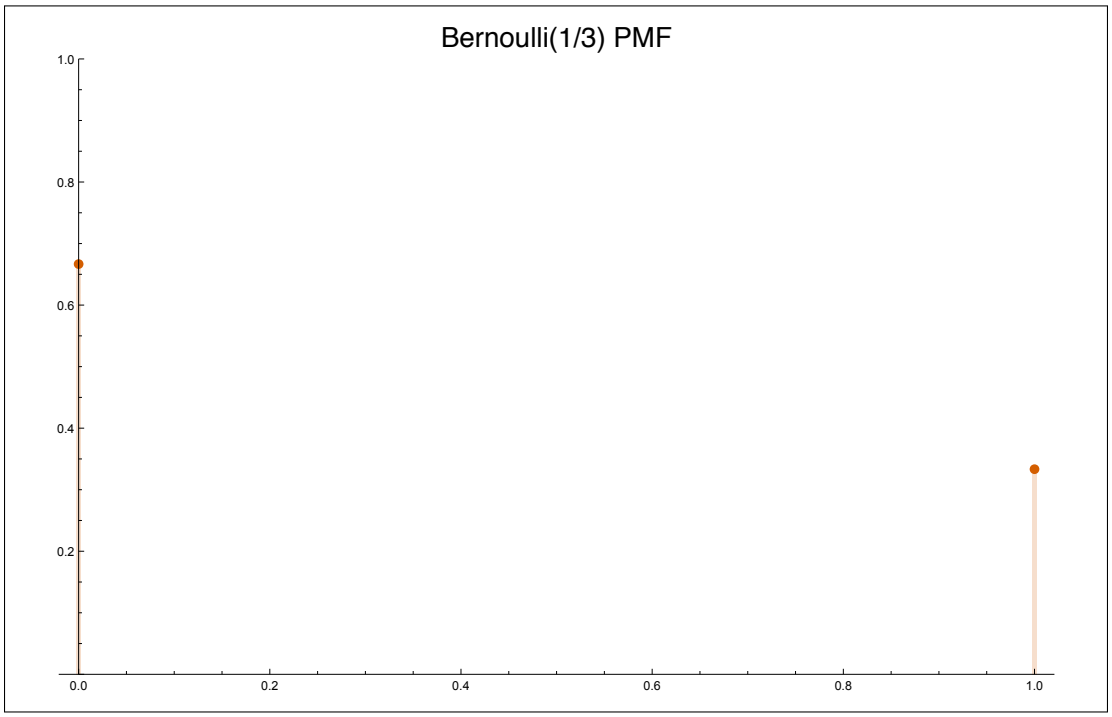


Figure 5.1. The Bernoulli pmf

### 5.7.2 Binomial random variables

The **Binomial$(n, p)$ distribution** is the distribution of the number $X$ of "successes" in $n$ independent Bernoulli$(p)$ trials. 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 pmfs are **unimodal**. The **mode** is the value where the pmf 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 pmf for $p = 0.5$ is symmetric about $pn$, the height of the mode is lower, and the pmf is more "spread out." The pmfs for $p = 0.2$ and $p = 0.8$ are mirror images, which should be obvious from the formula for the pmf.

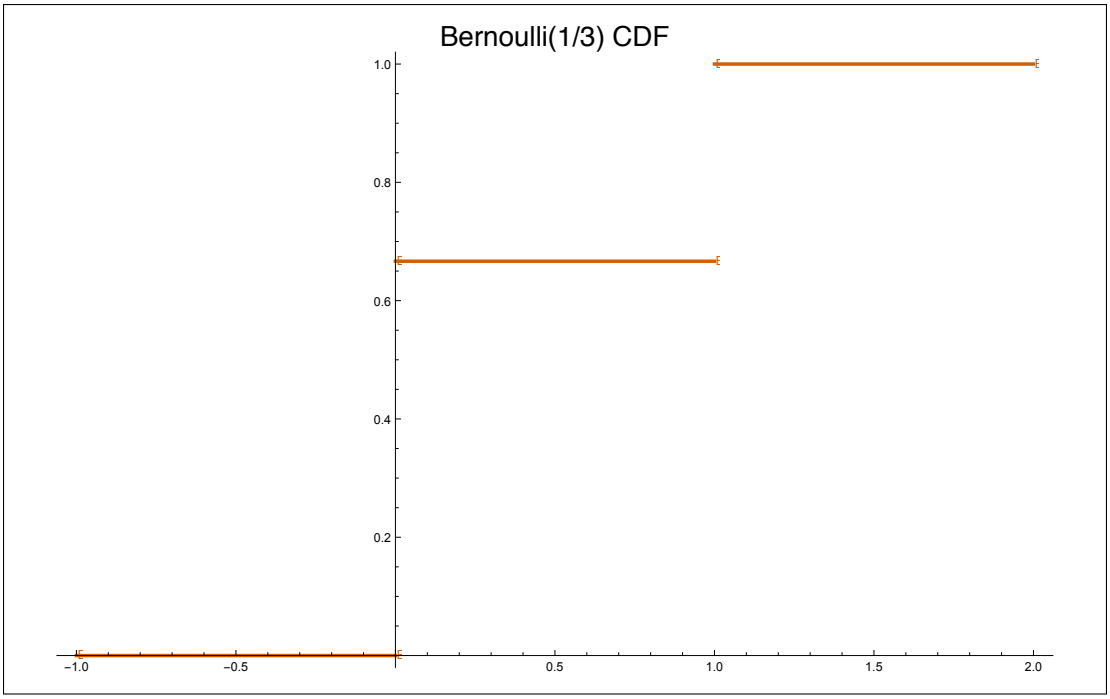


Figure 5.2. The Bernoulli cdf

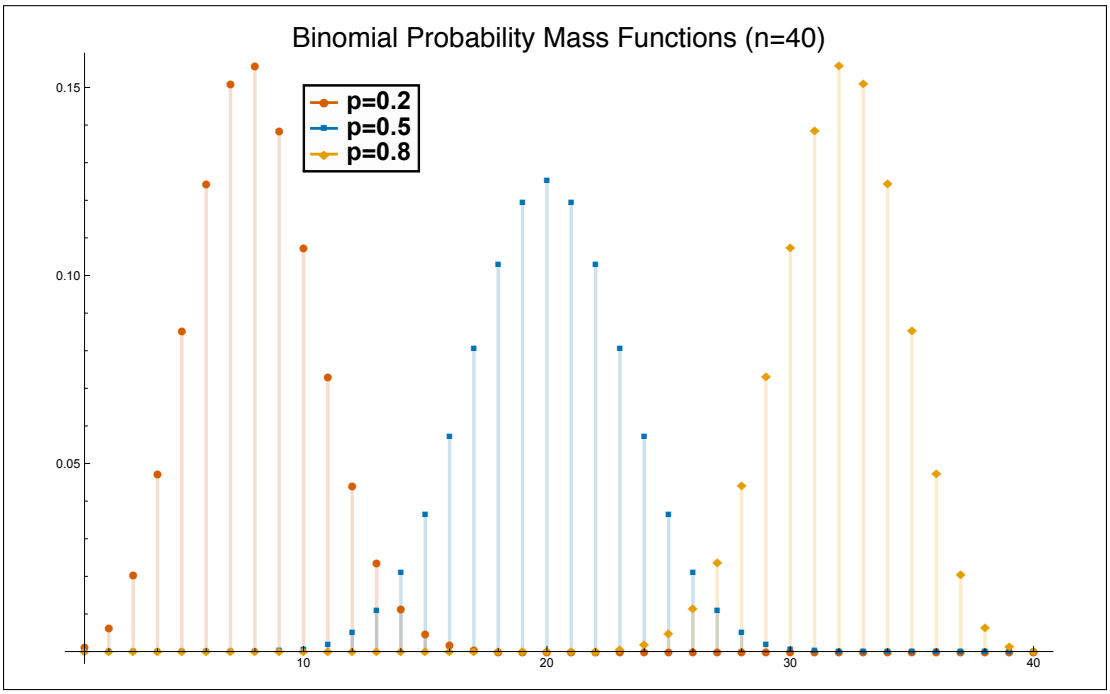


Figure 5.3. Binomial probability mass functions.

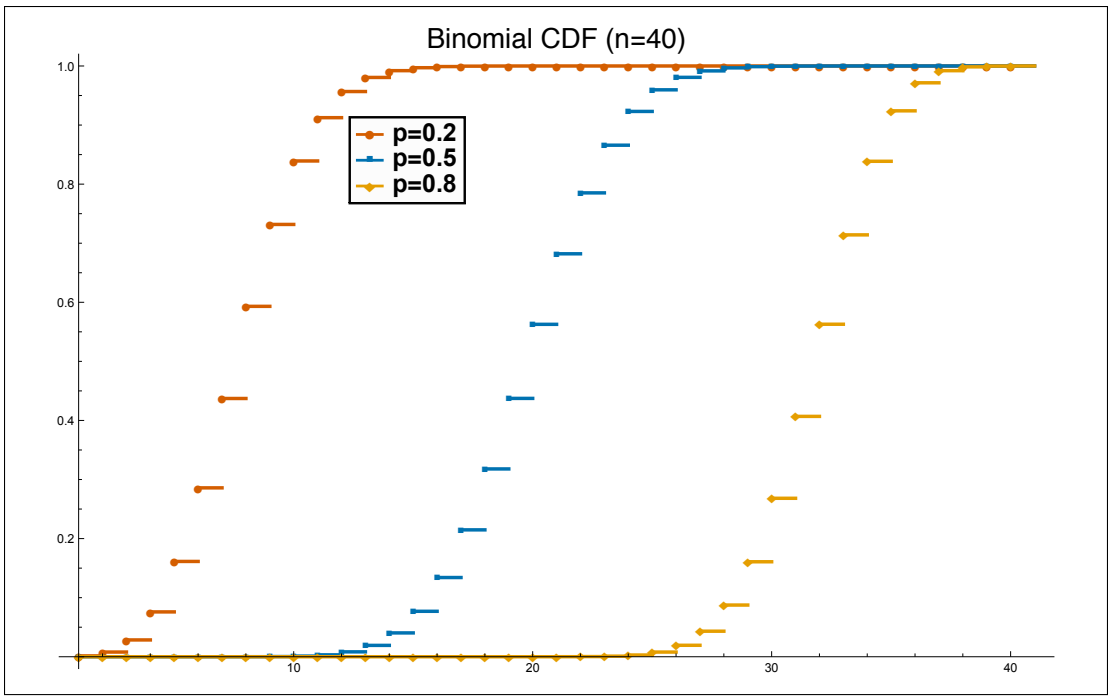


Figure 5.4. Binomial cumulative distribution functions.

## 5.8 ⋆ Stochastic dominance

Note: This material is in neither Pitman [5] nor Larsen–Marx [4].

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 Expectation

**Pitman [5]:** § 3.1
**Larsen–Marx [4]:** § 3.5

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

$$S = \{s_1, \ldots, s_n\},$$

and that the probability of each outcome is given by the probability measure $P$ on $S$. Suppose further that in outcome $s \in S$, you win $X(s)$. 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 $X$.

**Pitman [5]:** § 3.2

**5.9.1 Definition** *Let $S$ be a finite or denumerably infinite sample space and let $X$ be a random variable on $S$. The* ***expectation****, or* ***mathematical expectation****, or the* ***mean*** *of $X$ is defined to be*

$$\begin{aligned} \boldsymbol{E}\, X &= \sum_{s \in S} X(s)P(s), \\ &= \sum_{x \in \operatorname{range} X} xp(x), \end{aligned}$$

*where $p$ is the probability mass function; provided that in case $S$ is infinite, the series is absolutely convergent.*

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

The expectation of the indicator function $\mathbf{1}_A$ of an event $A$ is $P(A)$.

Note that

the expectation of $X$ is determined by its distribution on $\boldsymbol{R}$.

**N.B.** Note that $\boldsymbol{E}$ is an **operator** on the space of 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. There are a few expressions coming up where I may add extra parentheses for clarity.

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 $s_i$ once, we shall have won $\sum X(s)$. So the fair price per play should be $\sum X(s)/n = \boldsymbol{E}\,X$.

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

For a discrete random variable $X$ with values $x_1, x_2, \dots$ 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 $\mu$. From what I recall of Ph 1a, the total torque on the beam is

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

(provided $\sum_i p(x_i)x_i$ is absolutely convergent). Which value of $\mu$ 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[2] the expectation is also known as the **first moment** of the random variable's distribution.

It follows that

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

*Proof*: By definition,

$$\boldsymbol{E}\,X = \sum_i x_i p_i$$

and

$$\boldsymbol{E}\big(X - (\boldsymbol{E}\,X)\big) = \sum_i \big(x_i - (\boldsymbol{E}\,X)\big)p_i = \sum_i x_i p_i - (\boldsymbol{E}\,X)\sum_i p_i = \boldsymbol{E}\,X - \boldsymbol{E}\,X = 0,$$

provided the series $\sum_{i=1}^{\infty} x_i p_i$ is absolutely convergent. ∎

**5.9.3 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.

[2] According to my copy of the *OED* [7] the term "moment" comes from the Latin *momentum*, meaning "movement" or "moving force."

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

- The "fair price" of a gamble $X$.
- The location of the "center" 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 $P(E)$.

## 5.10 Expectation of a function of a discrete random variable

If $X$ is a discrete random variable on a probability space $(S, \mathcal{E}, P)$ and $g$ is a function from $\boldsymbol{R}$ to $\boldsymbol{R}$, then the composition $g \circ X$ is also a discrete random variable, so

$$\begin{aligned}\boldsymbol{E}(g \circ X) &= \sum_{s \in S} g\big(X(s)\big)P(s),\\ &= \sum_{x \in \operatorname{range} X} g(x)p(x)\end{aligned}$$

provided that in case $S$ is infinite, the series is absolutely convergent.

## 5.11 The St. Petersburg Paradox

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

**5.11.1 Example (The St. Petersburg Paradox)** (See also Larsen–Marx [4, Example 3.5.5, pp. 144–145].) Consider the following game: Toss a fair coin until the first Tails appears. If this happens on $n^{\text{th}}$ toss, you win $2^n$.

What is the expected value of this game?

$$\begin{aligned}\boldsymbol{E}\,\text{Value} &= \sum_{n=1}^{\infty} (\text{winnings if first Tails is on toss } n) \times \text{Prob}\,(\text{first Tails is on toss } n)\\ &= \sum_{n=1}^{\infty} 2^n \frac{1}{2^n}\\ &= \sum_{n=1}^{\infty} 1\\ &= \infty \text{ (?!)}\end{aligned}$$

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 $2^n < \infty$.

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 Wikipedia, "the paradox takes its name from its resolution by Daniel Bernoulli, one-time resident of the eponymous Russian city, who published his arguments in the *Commentaries of the Imperial Academy of Science of Saint Petersburg* (1738). However, the problem was invented by Daniel's brother Nicolas Bernoulli who first stated it in a letter to Pierre Raymond de Montmort on September 9, 1713."

**5.11.2 Remark** The expected length of a St. Petersburg game is

$$\sum_{k=1}^{\infty} k2^{-k} = 2.$$

For a derivation of the value of the series, see Supplement 1.

## 5.12 ⋆ Infinite Expectation and Nonexistent Expectation

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. For historical reasons, we shall denote the set of random variables on the sample space $(S, \mathcal{E}, P)$ that have a finite expectation by $L_1(P)$, or more simply by $L_1$.. In that case, its expectation is given by the formulas above. If $X$ is a nonnegative random variable, and the expectation formula gives an infinite value, we shall say the expectation is infinite, $\boldsymbol{E}\, X = \infty$. We may also have a random variable whose negative has infinite expectation, in which case we say its expectation is negative infinity, $-\infty$.

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 time period a step of size $\pm 1$ is taken with equal probability. We shall see in Lecture 16 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, 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 $n^{\text{th}}$ toss, you win $2^n$. if the die comes up odd, then if the first Tails happens on $n^{\text{th}}$ 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. In this case, we say that 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.

**5.12.1 Definition** *For a random variable $X$, define*

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

*the* ***positive part*** *of $X$ and the* ***negative part*** *of $X$, respectively. Note that*

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

*Let*

$$\begin{aligned} \boldsymbol{E}\, X &= \sum_{s \in S} X(s)P(s), \\ &= \sum_{x \in \text{range } X} xp(x), \end{aligned}$$

*where we allow the sum to diverge.*

- *If both $\boldsymbol{E}\, X^+ = \infty$ and $\boldsymbol{E}\, X^- = \infty$ then we say that the* ***expectation of $X$ does not exist****.*
- *If $\boldsymbol{E}\, X^+ = \infty$ and $\boldsymbol{E}\, X^- < \infty$, then the expectation of $X$ exists, but is infinite, $\boldsymbol{E}\, X = \infty$.*
- *If $\boldsymbol{E}\, X^+ < \infty$ and $\boldsymbol{E}\, X^- = \infty$, then the expectation of $X$ exists, but is negatively infinite, $\boldsymbol{E}\, X = -\infty$.*
- *Finally if both $\boldsymbol{E}\, X^+ < \infty$ and $\boldsymbol{E}\, X^- < \infty$, then the expectation of $X$ exists and is finite, and satisfies*

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

We'll come back to this in Section 6.6.

## 5.13 Independent random variables

**Pitman [5]:** pp. 151–154

**5.13.1 Definition** *$X$ and $Y$ are* ***independent random variables*** *if for every $B_1, B_2 \subset \boldsymbol{R}$,* [a]

$$P\left(X \in B_1 \textit{ and } Y \in B_2\right) = P\left(X \in B_1\right) \cdot P\left(Y \in B_2\right)$$

*More generally, a set $\mathcal{X}$ of random variables is* ***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 $\boldsymbol{R}$,*

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

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

**5.13.2 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 $X+Y$. Then $X$ and $Y$ are stochastically independent, $Y$ and $Z$ are stochastically independent, and $X$ and $Z$ are stochastically independent; but the set $X$, $Y$, $Z$ is *not* stochastically independent.

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

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

## Bibliography

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

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

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

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

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

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

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