Caltech Department of
Mathematics

Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2021

Lecture 2: Modeling Random Experiments

Relevant textbook passages:

Pitman [5]: Sections 1.3–1.4., pp. 26–46.

Larsen–Marx [4]: Sections 2.2–2.5, pp. 18–66.

2.1 Axioms for probability measures

Recall from last time that a **random experiment** is an experiment that may be conducted under seemingly identical conditions, yet give different results. Coin tossing is everyone's go-to example of a random experiment.

The way we model random experiments is through the use of probabilities. We start with the **sample space** Ω, the set of possible outcomes of the experiment, and consider **events**, which are subsets E of the sample space. (We let $\mathcal{F}$ denote the collection of events.)

2.1.1 Definition *A* ***probability measure*** *P or* ***probability distribution*** *attaches to each event E a number between 0 and 1 (inclusive) so as to obey the following* ***axioms of probability****:*

Normalization: *$P(\varnothing) = 0$; and $P(\Omega) = 1$.*

Nonnegativity: *For each event E, we have $P(E) \geqslant 0$.*

Additivity: *If $EF = \varnothing$, then $P(\cup F) = P(E) + P(F)$.*

Note that while the domain of P is technically $\mathcal{F}$, the set of events, that is $P \colon \mathcal{F} \to [0,1]$, we may also refer to P as a probability (measure) on Ω, the set of realizations.

2.1.2 Remark To reduce the visual clutter created by layers of delimiters in our notation, we may omit some of them simply write something like $P(f(\omega) = 1)$ or $P\{\omega \in \Omega : f(\omega) = 1\}$ instead of $P\big(\{\omega \in \Omega : f(\omega) = 1\}\big)$ and we may write $P(\omega)$ instead of $P\big(\{\omega\}\big)$. You will come to appreciate this.

Most probabilists also require the following stronger property, called **countable additivity**.

2.1.3 Definition *A probability measure P is* ***countably additive*** *if for every pairwise disjoint infinite sequence of events $E_1, E_2, \dots$,*

$$P\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \sum_{i=1}^{\infty} P(E_i).$$

2.1.4 Remark If the sample space Ω is finite, it has only finitely many subsets, so the only way an infinite sequence $E_1, E_2, \dots$ of events can be pairwise disjoint is that all but finitely many of the events E_i are empty. In this case, since $P(\varnothing) = 0$, the infinite series $\sum_{i=1}^{\infty} P(E_i)$ reduces to a finite sum. In other words, for a finite sample space, finite additivity guarantees countable additivity. (Cf. Section 2.2.1, item 4.)

You need to take an advanced analysis course to understand that for infinite sample spaces, there can be probability measures that are additive, but not countably additive. So don't worry too much about it.

From here on out, we shall assume that all our probability measures are countably additive.

2.1.1 The special probability values zero and one

It is important to distinguish between an event that cannot happen and one that has probability zero. Any decent model of a random experiment should assign probability zero to an event that cannot happen, but an event may still be possible and yet have probability zero. We shall see presently that the probability of an infinite sequence of nothing but Tails has probability zero (in the standard model of coin tossing), but that does not mean that it is impossible. As far as I know, such a sequence would not violate the laws of physics, but I should check with the Physics Police on that. (Actually if coin tossing takes time, and the life of the universe is finite we can never carry out the experiment anyhow.)

Likewise an event can have probability one, and yet may not necessarily occur. At least that is the interpretation I shall take in this course. As I mentioned earlier, I do not want to become bogged down in metaphysics. But probabilists have coined a phrase to describe events of probability one.

2.1.5 Definition *If $P(E) = 1$, then we say that that E occurs* ***almost surely****, abbreviated* ***E a.s.*** *An event of probaility zero is sometimes called a* **null event**.

You might ask why be wishy-washy and add the "almost?" We shall have more to say about this in Section 6.3.

2.2 Probability spaces

Our complete formal model of a **random experiment** is what we call a probability space.

2.2.1 Definition *A* ***probability space*** *is a triple $(\Omega, \mathcal{F}, P)$, where Ω is a nonempty set, the* ***sample space*** *or* ***outcome space*** *of the experiment, $\mathcal{F}$ is the set of* ***events****, which is a σ-algebra of subsets of Ω, and P is a countably additive* **probability** *measure on $\mathcal{F}$.*

2.2.1 Elementary Probability Identities

There are some elementary properties of probability measures that follow in a straightforward way from the axioms. Here are a few together with their proofs (derivations, if you prefer).

Larsen–Marx [4]: § 2.3

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

Proof: Since E and E^c are disjoint, and $E \cup E^c = \Omega$, additivity gives us $P(E) + P(E^c) = P(\Omega)$. Normalization gives us $P(\Omega) = 1$, and the desired expression is just a rearrangement. ∎

2. If $F \subset E$, then

$$P(E \setminus F) = P(E) - P(F)$$

Proof: Recall the definition of $E \setminus F$ as the set of points in E that are not in F, that is EF^{c}. Thus $E = F \cup (E \setminus F)$ where F and $E \setminus F$ are disjoint. By additivity then $P(E) = P(F) + P(E \setminus F)$, and the desired expression is just a rearrangement. ∎

3. **Monotonicity**: If $F \subset E$, then

$$P(F) \leqslant P(E)$$

Proof: We just saw that $P(E) = P(F) + P(E \setminus F)$, and nonnegativity implies $P(E \setminus F) \geqslant 0$. Therefore $P(E) \geqslant P(F)$. ∎

This has two useful special cases:

a. If $P(E) = 0$ and $F \subset E$, then $P(F) = 0$.

b. If $P(E) = 1$ and $E \subset F$, then $P(F) = 1$.

4. **Finite additivity**: If $E_1, \dots, E_n$ is a finite sequence of **pairwise disjoint** events, that is, $i \neq j \implies E_i E_j = \varnothing$, then

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i).$$

Proof: We shall prove this by induction on n. Let $\mathbb{P}(n)$ stand for the above equality for n pairwise disjoint sets. Then $\mathbb{P}(2)$ is just Additivity. Assume $\mathbb{P}(n-1)$. Write

$$\bigcup_{i=1}^{n} E_i = \underbrace{\bigcup_{i=1}^{n-1} E_i}_{=B} \cup E_n$$

Then $BE_n = \varnothing$, so

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} E_i\right) &= P\left(B \cup E_n\right) \\ &= P(B) + P(E_n) \qquad \text{by Additivity} \\ &= \sum_{i=1}^{n-1} P(E_i) + P(E_n) \quad \text{by } \mathbb{P}(n-1) \\ &= \sum_{i=1}^{n} P(E_i). \end{aligned}$$

∎

Boole's inequality

The next result is obvious, but it can be derived from our axioms.

2.2.2 Boole's Inequality *Even if events $E_1, \dots, E_n$ are not pairwise disjoint, we may still conclude that*

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leqslant \sum_{i=1}^{n} P(E_i).$$

Before I demonstrate how to prove Boole's Inequality, let me describe a "trick" for "disjunctifying" a sequence of sets.

2.2.3 Lemma *Let $E_1, \dots, E_n, \dots$ be a sequence (finite or infinite) of events. Then there is a sequence $A_1, \dots, A_n, \dots$ of events such that:*

- *For each n, $A_n \subset E_n$.*
- *The A_i's are pairwise disjoint. That is, if $i \neq j$, then $A_i A_j = \varnothing$.*
- *For each n,*
$$\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} E_i.$$
- *The last fact implies that*
$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} E_i.$$

Proof: Define

$$A_1 = E_1,$$

and for $n > 1$ define

$$A_n = E_n \setminus (E_1 \cup \dots \cup E_{n-1}) = E_n \setminus (A_1 \cup \dots \cup A_{n-1}).$$

I leave it to you to prove that the sequence $A_1, \dots, A_n, \dots$ has the desired properties. See Figure 2.1 for an example with three events. ∎

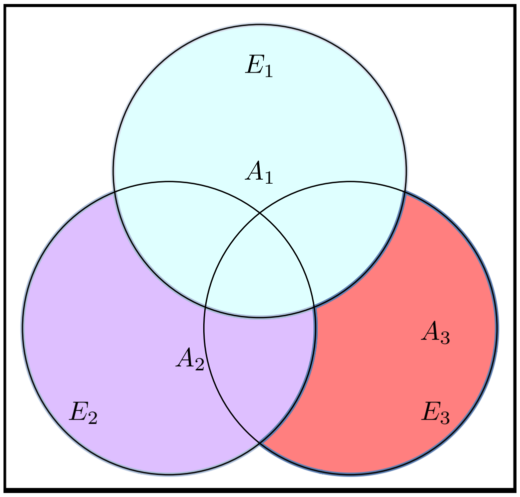


Figure 2.1. The E_i events are the discs, and the three colored regions represent the events $A_1 = E_1$, $A_2 = E_2 \setminus E_1$, and $A_3 = E_3 \setminus (E_1 \cup E_2)$.

Proof of Boole's Inequality: Let A_i be a sequence of pairwise disjoint events satisfying the conclusion of Lemma 2.2.3. That is, each $A_i \subset E_i$ and $\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} E_i$. Then

$$P\left(\bigcup_{i=1}^{n} E_i\right) = P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \leqslant \sum_{i=1}^{n} P(E_i),$$

where the second equality follows from the fact that the A_i's are pairwise disjoint and finite additivity, and the inequality follows from $P(A_i) \leqslant P(E_i)$ (monotonicity) for each i. ∎

Bonferroni's inequality

Boole's inequality provides an upper bound on the probability of a union of not necessarily disjoint events. Bonferroni's inequality flips this over and gives a lower bound on the probability of an intersection of events. Note that the right-hand side of Bonferroni's inequality can be negative, which gives a valid, but uninformative, bound.

2.2.4 Bonferroni's Inequality *For a finite sequence of events $E_1, \dots, E_n$,*

$$P\left(\bigcap_{i=1}^{n} E_i\right) \geqslant \sum_{i=1}^{n} P(E_i) - (n-1).$$

I'll leave the proof as an exercise. But here's a big hint. Use de Morgan's Laws (specifically $(\cap E_i)^{\mathrm{c}} = \cup E_i^{\mathrm{c}}$), the fact that $P(E^{\mathrm{c}}) = 1 - P(E)$, and Boole's inequality.

Here are two handy consequences of Boole's Law and Bonferroni's Inequality.

2.2.5 Proposition *If $P(E_1) = \cdots = P(E_n) = 0$, then $P\big(\bigcup_{i=1}^{n} E_i\big) = 0$.*

If $P(E_1) = \cdots = P(E_n) = 1$, then $P\big(\cap_{i=1}^{n} E_i\big) = 1$.

2.2.2 Digression on terminology: Odds

Indulge me while I vent about one of my pet peeves. I find it exasperating that even generally linguistically reliable sources, such as *The New York Times* or *The Economist*, confuse probabilities and odds.

Pitman [5]: pp. 6–8

2.2.6 Definition *The* ***odds against the event E*** *is the ratio*

$$\frac{P(E^{\mathrm{c}})}{P(E)}.$$

The ***odds in favor of the event E*** *is the ratio*

$$\frac{P(E)}{P(E^{\mathrm{c}})}.$$

That is, odds are a ratio of probabilities of an event and its complement, not the probability of an event.[1] It is customary to say something like "the odds against the event E are $P(E^{\mathrm{c}}) : P(E)$" as a ratio of integers. That is, we typically say "3:2", pronounced "3 to 2," instead of "$1\frac{1}{2}$".

Unfortunately, you often run across statements such as, "The Upshot puts odds of a Republican takeover of the Senate at 74 percent," from *The New York Times*, when they mean that the odds are roughly 3:1 in favor of a takeover; or from *The Economist*, "the Federal Reserve Bank of New York puts the odds that a recession might begin within the next 12 months at

[1] You may hove noticed that I used the singular verb "is" in the definition above and the plural verb "are" in the sentences preceding and following this note. The word "odds" is both plural and singular. When in doubt, you can always say "the odds ratio is ..."

over 37%," when they mean the probability is 37%, or the odds are roughly 3:5 in favor of a recession.

This is to be distinguished from the **payoff odds**. The payoff odds are the ratio of the amount won to the amount wagered for a simple **bet**. For instance, in roulette, if you bet $1 on 2 and 2 comes up, you get a payoff of $35 (in addition to the return of your dollar), so the payoff odds are "35 to 1." But (assuming that all numbers on a roulette wheel are equally likely) the odds against 2 are 37 to 1 since a roulette wheel has the "numbers" 0 and 00 in addition to the numbers 1 through 36.[2] [3] [4] (Pitman [5, p. 7] describes the outcomes and bets for a roulette wheel.) In poker, the **pot odds**, which is the ratio of the size of the pot to the amount of the bet, are relevant to deciding whether to call a bet or fold.

2.3 Additivity and the Inclusion–Exclusion Principle

The **Inclusion–Exclusion Principle** describes the full power of additivity of probability measures when applied to unions of not necessarily pairwise disjoint sets. Early on, we expect small children to understand the relation between sets and their cardinality—If Alex has three apples and Blair has two apples, then how many apples do they have together? The implicit assumption is that the two sets of apples are disjoint (since they belong to different children), then the measure (count) of the union is the sum of the counts. But what if Alex and Blair own some of their apples in common?

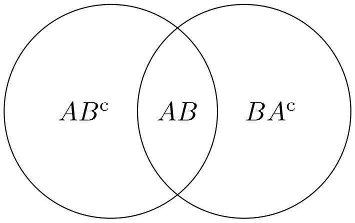
Pitman [5]: p. 22

2.3.1 Proposition (Inclusion–Exclusion Principle, I) *Even if $AB \neq \varnothing$,*

$$P(A \cup B) = P(A) + P(B) - P(AB).$$

Proof: Now

$$A \cup B = (AB^{c}) \cup (AB) \cup (A^{c}B).$$



The three sets in the union on the right-hand side are pairwise disjoint:

$$(AB^{c})(AB) = \varnothing$$
$$(AB)(A^{c}B) = \varnothing$$
$$(AB^{c})(A^{c}B) = \varnothing.$$

Therefore, by finite additivity,

$$P(A \cup B) = P(AB^{c}) + P(AB) + P(A^{c}B).$$

Now also by additivity,

$$P(A) = P(AB^{c}) + P(AB)$$
$$P(B) = P(BA^{c}) + P(AB).$$

[2] Actually, there are (at least) two kinds of roulette wheels. In Las Vegas, roulette wheels have 0 and 00, but in Monte Carlo, the 00 is missing.

[3] The term roulette wheel is a pleonasm, since *roulette* is French for "little wheel."

[4] The word "pleonasm" is one of my favorites. Look it up.

So, adding and regrouping,

$$P(A) + P(B) = \underbrace{P(AB^{c}) + P(AB) + P(BA^{c})}\,+P(AB)$$

$$= P(A \cup B) + P(AB).$$

This implies

$$P(A \cup B) = P(A) + P(B) - P(AB).$$

∎

Additionally,

$$\boxed{\begin{aligned} P(A \cup B \cup C) = {} & P(A) + P(B) + P(C) \\ & - P(AB) - P(AC) - P(BC) \\ & + P(ABC). \end{aligned}}$$

To see this refer to Figure 2.2. The events A, B, and C are represented by the three circles.

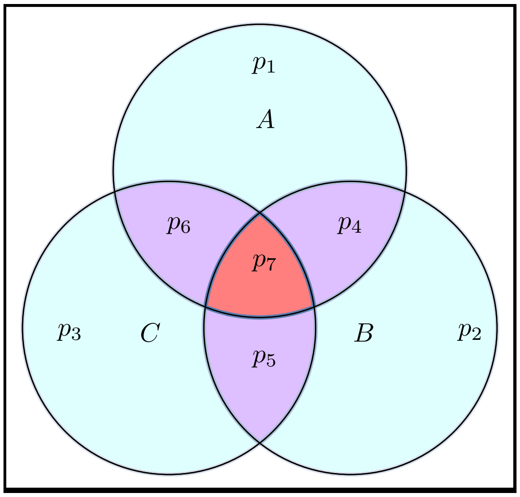


Figure 2.2. Inclusion/Exclusion for three sets.

The probability of each shaded region is designated by p_i, $i = 1, \dots, 7$. Observe that

$$\begin{gathered} P(A \cup B \cup C) = p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7 \\ P(A) = p_1 + p_4 + p_6 + p_7 \\ P(B) = p_2 + p_4 + p_5 + p_7 \\ P(C) = p_3 + p_5 + p_6 + p_7 \\ P(AB) = p_4 + p_7 \\ P(AC) = p_6 + p_7 \\ P(BC) = p_5 + p_7 \\ P(ABC) = p_7. \end{gathered}$$

Thus

$$\begin{aligned} P(A)+P(B)+P(C) &= p_1+p_2+p_3+2p_4+2p_5+2p_6+3p_7 \\ P(AB)+P(AC)+P(BC) &= p_4+p_5+p_6+3p_7. \end{aligned}$$

So

$$\begin{aligned} &\big[P(A)+P(B)+P(C)\big]-\big[P(AB)+P(AC)+P(BC)\big]+P(ABC) \\ &\qquad\qquad\qquad = p_1+p_2+p_3+p_4+p_5+p_6+p_7 = P(A\cup B\cup C). \end{aligned}$$

$\blacksquare$

The general version of the Inclusion–Exclusion Principle may be found in Pitman [5], Exercise 1.3.12, p. 31.

2.3.2 Proposition (General Inclusion–Exclusion Principle)

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} E_i\right) = &\sum_i P(E_i) \\ &-\sum_{i<j} P(E_iE_j) \\ &+\sum_{i<j<k} P(E_iE_jE_k) \\ &\;\vdots \\ &+(-1)^{n+1}P(E_1E_2\cdots E_n) \\ =&\sum_{k=1}^{n}\sum_{i_1<\cdots<i_k}(-1)^{k+1}E_{i_1}\cdots E_{i_k}. \end{aligned}$$

(Recall that intersection is denoted by placing sets next to each other. Note that the sign preceding a sum with the intersection of m sets is $(-1)^{m+1}$. The reason for summing over increasing indices is to avoid double counting.)

Note that if the sets are pairwise disjoint, the intersections above are all empty and so have probability zero, and this reduces to finite additivity.

Here is a sketch of how to prove this result using induction on n, but after we have learned about the expectation of random variables, the proof will be easier to follow. See Theorem 8.3.3 below.

Proof: Here is the outline of a proof by induction on n. Given $E_1,\dots,E_{n+1}$, let $A=\bigcup_{i=1}^{n-1} E_i$ and $B=E_n\cup E_{n+1}$. Then $P(A\cup B)=P(A)+P(B)-P(AB)$. Expand

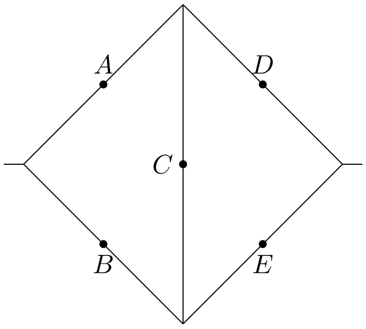
$$\begin{aligned} P(A) = P\big(\bigcup_{i=1}^{n-1} E_i\big) &= \sum_{k=1}^{n-1}\sum_{i_1<\cdots<i_k}(-1)^{k+1}E_{i_1}\cdots E_{i_k}, \\ P(B) = (E_n\cup E_{n+1}) &= P(E_n)+P(E_{n+1})-P(E_nE_{n+1}), \end{aligned}$$

and

$$\begin{aligned} P(AB) &= P\big(\big(\bigcup_{i=1}^{n-1} E_iE_n\big)\cup\big(\bigcup_{i=1}^{n-1} E_iE_{n+1}\big)\big) \\ &\qquad\qquad = P\big(\bigcup_{i=1}^{n-1} E_iE_n\big)+P\big(\bigcup_{i=1}^{n-1} E_iE_{n+1}\big)-P(\cap_{i=1}^{n-1}E_iE_nE_{n+1}). \end{aligned}$$

Then expand everything and combine like terms, and you should get the desired result. $\blacksquare$

2.3.3 Example (An application) Consider the electric circuit described by the diagram below.



There are switches at the dots labeled A, B, C, D, and E. Gremlins randomly open and close the switches. Let A, B, C, D, and E denote the event that the corresponding switch is closed, and let $a, \dots, e$ be the corresponding probabilities. Assume that the gremlins make these events stochastically independent. The question is

What is the probability that a current can flow though the network?

There are four events in which the current can flow:

$$F_1 = AD, \quad F_2 = BE, \quad F_3 = ACE, \quad F_4 = BCD,$$

so what we need is the probability of this union. But these events are not disjoint, so we need the Inclusion/Exclusion Principle to get

$$\begin{aligned} &P(F_1 \cup F_2 \cup F_3 \cup F_4) = \\ &\qquad P(F_1) + P(F_2) + P(F_3) + P(F_4) \\ &\qquad -P(F_1F_2) - P(F_1F_3) - P(F_1F_4) - P(F_2F_3) - P(F_2F_4) - P(F_3F_4) \\ &\qquad +P(F_1F_2F_3) + P(F_1F_2F_4) + P(F_1F_3F_4) + P(F_2F_3F_4) \\ &\qquad -P(F_1F_2F_3F_4). \end{aligned}$$

Since $A, \dots, E$ are stochastically independent events, we have $P(F_1F_2) = P(ADBE) = adbe$, $P(F_1F_3) = P(ADACE) = P(ACDE) = acde$ (don't be fooled into thinking that $P(ADACE) = a^2cde$, it's $acde$), etc. By my reckoning, the probability that current flows is

$$ad + be + ace + bcd - abcd - abce - abde - aced - bcde + 2abcde.$$

It's not pretty, but it is useful. Let me know if it's wrong.

Note that in this problem the sample space can be identified with $\{0,1\}^5$, which has 32 points, corresponding to the possible configurations of the five switches. (E.g., the point $(1,1,0,1,0)$ corresponds to switches at the dots A, B, and D being closed and the others open.) These points need not be equally likely unless the probability for each switch is 1/2. The event we call A is composed of sixteen points of the sample space, those whose first coordinate is 1. The event $F_1 = AD$ is composed of eight points in the sample space, while $F_3 = ACE$ contains four points. □

2.4 ⋆ More on countable additivity

The next results may seem theoretical and of no practical relevance, but they are crucial to understanding the properties of cumulative distribution functions.

2.4.1 Definition *A sequence $E_1, \dots, E_n \dots$ of events is* ***decreasing****, written $E_n \downarrow$, if*

$$E_1 \supset E_2 \supset \cdots \supset E_n \supset \cdots .$$

In this case, letting $E = \bigcap\limits_n E_n$ we may write $E_n \downarrow E$ or $\lim_{n\to\infty} E_n = E$. A sequence $E_1, \dots, E_n \dots$ of events is ***increasing****, written $E_n \uparrow$, if*

$$E_1 \subset E_2 \subset \cdots \subset E_n \subset \cdots .$$

In this case, letting $E = \bigcup\limits_n E_n$ we may write $E_n \uparrow E$ or $\lim_{n\to\infty} E_n = E$.

2.4.2 Definition (Continuity of set functions) *A set function P is* ***continuous from above*** *if $E_n \downarrow E$ implies $\lim_n P(E_n) = P(E)$. A set function P is* ***continuous from below*** *if $E_n \uparrow E$ implies $\lim_n P(E_n) = P(E)$.*

2.4.3 Proposition (Continuity and countable additivity) *Let Ω be a set and $\mathcal{F}$ be a σ-algebra of subsets of Ω. Let P be an additive probability on $\mathcal{F}$. Then the following are equivalent.*

1. *P is countably additive.*
2. *P is continuous from above.*
3. *P is continuous from below.*

Proof: (1) $\implies$ (2): Assume that P is countably additive, and that $E_n \downarrow E$. We need to show that $P(E) = \lim_{n\to\infty} P(E_n)$. To do so, we first construct a sequence $A_1, A_2, \dots$ of pairwise disjoint events such that $\bigcup\limits_{n=1}^{\infty} A_n = E^{\mathrm{c}}$. Start by setting $A_1 = \Omega \setminus E_1$, and for $n > 1$ recursively define $A_n = E_n \setminus E_{n-1}$. (It helps to draw a Venn diagram.) Since the E_n's are decreasing, each A_n is disjoint from its predecessors, and $\bigcup\limits_{i=1}^{n} A_i = E_n^{\mathrm{c}}$. Thus $\bigcup\limits_{n=1}^{\infty} A_n = E^{\mathrm{c}}$. Then

$$P\Bigl(\bigcup_{i=1}^{n} A_i\Bigr) = P(E_n^{\mathrm{c}}) = 1 - P(E_n),$$

and letting $n \to \infty$ we get

$$1 - P(E) = P(E^{\mathrm{c}}) = P\Bigl(\bigcup_{i=1}^{\infty} A_i\Bigr) = \sum_{i=1}^{\infty} P(A_i) = \lim_{n\to\infty} \sum_{i=1}^{n} P(A_i) = 1 - \lim_{n\to\infty} P(E_n),$$

and the conclusion follows.

(2) $\implies$ (3): Assume that P is continuous from above, and let $A_n \uparrow A$. Let $E_n = \bigcup\limits_{i=n+1}^{\infty} A_i$. Then $E_n \downarrow \varnothing$, so by continuity from above, $\lim_n P(E_n) = P(\varnothing) = 0$. Now

$$A = \bigcup_{i=1}^{\infty} A_i = A_n \cup E_n$$

for each n. Thus since P is finitely additive

$$P\bigl(\bigcup_{i=1}^{\infty} A_i\bigr) = P(A_n \bigcup E_n) = P(A_n) + P(E_n).$$

But the left-hand side is independent of n, and so letting $n \to \infty$ we get

$$P(A) = \lim_n P(A_n) + \underbrace{\lim_n P(E_n)}_{=0}.$$

(3) $\implies$ (1): Assume that P is continuous from below, and let A_n be a sequence of pairwise disjoint events. Define $E_n = A_1 \bigcup \cdots \bigcup A_n$, so that $E_n \uparrow \bigcup_{n=1}^{\infty} A_n$, so by continuity form below,

$$P\big(\bigcup_{i=1}^{\infty} A_i\big) = \lim_{n\to\infty} P(E_n).$$

But since P is finitely additive

$$P(E_n) = \sum_{i=1}^{n} P(A_i)$$

so

$$P\big(\bigcup_{i=1}^{\infty} A_i\big) = \lim_{n\to\infty} P(E_n) = \lim_{n\to\infty} \sum_{i=1}^{n} P(A_i) = \sum_{i=1}^{\infty} P(A_i).$$

That is, P is countably additive. ∎

We can use this to extend Proposition 2.2.5.

2.4.4 Proposition *For a countably additive probability P, if $P(E_1) = P(E_2) = \cdots = 0$, then $P\big(\bigcup_{i=1}^{\infty} E_i\big) = 0$.*

Also if $P(E_1) = P(E_2) = \cdots = 1$, then $P\big(\cap_{i=1}^{\infty} E_i\big) = 1$.

2.5 Data: Random variables and random vectors

For some reason, your textbooks postpone the definition of random variables, even though they are fundamental concepts.

A **random variable** is a numerical measurement of the outcome of a random experiment. Values of random variables constitute the **data** from a random experiment.

2.5.1 Example (Some random variables) Here are some examples of random variables.

- The random experiment is to roll two dice, so the sample space is the set of ordered pairs of integers from 1 through 6. The sum of these to numbers is a random variable. The pair itself is a random vector. The difference of the numbers is another random variable.

- The experiment is to roll two dice repeatedly until boxcars (a pair of sixes) appear. The number of rolls is a random variable, which may take on the value ∞ if boxcars never appear. (This is an idealization of course.)

- The random experiment takes a sample of blood and smears it on a microscope slide with a little rectangle marked on it. The number of platelets lying in the rectangle is a random variable.

- The experiment is to record all the earthquakes in Southern California. The number of magnitude 5 + earthquakes in a year is a random variable.

- A letter is drawn at random from the alphabet. If we assign a number to each letter, then that number is a random variable. But unlike the cases above it does not make sense to take the results of two such experiments and add them. What letter is the sum of 'a' and 'b'? In such cases, where the result of the experiment is **categorical** and not inherently numeric, it may make more sense to take the outcome to be a random vector, indexed by the categories. This interpretation is often used in communication theory by electrical engineers, e.g., Robert Gray [3] or Thomas Cover and Joy Thomas [2].

- An experiment by Rutherford, Chadwick, and Ellis counted the number of α-particles emitted by a radioactive sample for consecutive 7.5-second time intervals. Each count is a random variable.

□

Being numerical, we can add random variables, take ratios, etc., to get new random variables. But to understand how these are related we have to go back to our formal model of random experiments as probability spaces, and define random variables in terms of a probability space.

2.5.2 Definition *A* ***random variable*** *on a probability space* $(\Omega, \mathcal{F}, P)$ *is an (extended)*[a] *real-valued* $\mathcal{F}$*-measurable function on* Ω.[b] [c]

A ***random vector*** *is simply a finite-dimensional vector (ordered list) of random variables.*

[a] The extended real numbers include two additional symbols, ∞ and $-\infty$. We'll have more to say about them later.

[b] Definition 2.10.1 below defines a function to $\mathcal{F}$-measurable if for every interval $I \subset \boldsymbol{R}$ the inverse image of I is an event in $\mathcal{F}$.

[c] Note that when the collection $\mathcal{F}$ of events consists of all subsets of Ω, then the requirement that inverse images of intervals be events is automatically satisfied.

So a random variable is not a variable in the usual sense of the word "variable" in either mathematics or computer programming. A random variable is simply an (extended) real-valued function. 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*

$$X \colon \Omega \to \boldsymbol{R} \quad \text{such that for each interval } I \subset \boldsymbol{R}, \quad \{\omega \in \Omega : X(\omega) \in I\} \in \mathcal{F}.$$

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

$$(X \in I) \text{ means } \{\omega \in \Omega : X(\omega) \in I\}.$$

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

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

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

is a random variable.

Other notations for an indicator function include I and double brackets, which are usually used with more complicated descriptions of events, such as

$$I\{X \in B\} = [\![X \in B]\!] = \mathbf{1}_{(X \in B)}.$$

That is,

$$I\{X \in B\}(\omega) = \begin{cases} 1 & X(\omega) \in B \\ = [\![X \in B]\!](\omega) & \\ 0 & X(\omega) \notin B. \end{cases}$$

A random variable X, is a **mapping** from the sample space Ω to the real numbers, that is, X maps each point $\omega \in \Omega$ to a real number $X(\omega)$. The function X is different from its value $X(\omega)$ at the point ω, which is simply a real number. The value $X(\omega)$ is frequently referred to as a **realization** of the random variable X. A realization is just the value that X takes on for some outcome ω in the sample space.

In these notes I will try to adhere to the convention that random variables (functions) are denoted by upper case letters, e.g., X, Y, etc., while realizations (values of the functions) are denoted by lower case letters, e.g., x, y, etc.

2.5.1 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

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

Similarly, a random m-vector $\boldsymbol{X} = (X_1, \dots, X_m)$ induces a probability measure or distribution on $\boldsymbol{R}^{\mathrm{m}}$ by

$$P_{\boldsymbol{X}}(I_1 \times \cdots \times I_m) = P\big(\{\omega \in \Omega : X_i(\omega) \in I_i,\ i = 1, \dots, m\}\big).$$

We shall discuss this further in Section 5.3.

2.6 Independent events

Larsen–Marx [4]: § 2.5
Pitman [5]: § 1.4

2.6.1 Definition *Events E and F are* ***(stochastically) independent*** *if*

$$P(EF) = P(E) \cdot P(F).$$

More generally, a set $\mathcal{A}$ of events is ***mutually independent*** *if for every finite subcollection $E_1, \dots, E_n$ of events in $\mathcal{A}$, we have*

$$P(E_1E_2 \cdots E_n) = P(E_1) \cdot P(E_2) \cdots P(E_n).$$

Whether or not two events are independent is determined by the probability measure, but note that for any event F, we have that F and Ω are independent and also that F and $\varnothing$ are independent.

2.6.2 Exercise Here's a little exercise that generalizes the last remark. Show that if E is any event and F is an event with $P(F) = 0$, then E and F are independent. Also show that if E is any event and F is an event with $P(F) = 1$, then E and F are independent. □

2.6.3 Lemma *If E and F are independent, then E and F^{c} are independent; and E^{c} and F^{c} are independent; and E^{c} and F are independent.*

Proof: It suffices to prove that if E and F are independent, then E^{c} and F are independent. The other conclusions follow by symmetry. So write

$$F = (EF) \cup (E^{\mathrm{c}}F),$$

so by additivity

$$P(F) = P(EF) + P(E^{c}F) = P(E)P(F) + P(E^{c}F),$$

where the second equality follows from the independence of E and F. Now solve for $P(E^{c}F)$ to get

$$P(E^{c}F) = \big(1 - P(E)\big)P(F) = P(E^{c})P(F).$$

But this is just the definition of independence of E^{c} and F. ■

2.7 Repeated experiments and product spaces

One of the chief uses of the theory of probability is to understand long-run frequencies of outcomes of experiments. If Ω is the sample space of a random experiment and $\mathcal{F}$ is its set of events, and we repeat the experiment, we have a **compound experiment** or **joint experiment** whose sample space is the Cartesian product $\Omega^2 = \Omega \times \Omega$, the set of ordered pairs of outcomes. Similarly, the sample space for n repetitions of the experiments is $\Omega^n = \underbrace{\Omega \times \cdots \times \Omega}_{n \text{ copies of } \Omega}$. The individual experiments in such a sequence are called **trials**.

The set of events for the compound experiment should certainly include sets of the form set of the form

$$E = E_1 \times \cdots \times E_n,$$

where each E_i is an event in $\mathcal{F}$, the common set of events for a single experiment is an event for the repeated experiment. Such a set is called a **rectangle**. (Geometrical rectangles are products of intervals.) But there are also non-rectangular events, for instance, the event that the outcomes are the same in each trial,

$$\{(s_1, \dots, s_n) : s_1 = s_2 = \cdots = s_n\}.$$

We want the set of events in the compound experiment to be an algebra. The smallest algebra that includes all the rectangles is called the **product algebra** and is denoted $\mathcal{F}^n$.

For example consider rolling a die, where every subset of $\{1, \dots, 6\}$ is an event in $\mathcal{F}$. The sample space for the repeated experiment is the set of ordered pairs of the numbers one through six. The event "both rolls give the same result" is not a rectangular event, but it is the union of finitely many rectangles: $= \bigcup_{i=1}^{6} \big(\{i\} \times \{i\}\big)$, and so is an event in the product algebra. So is the complementary event, "the rolls are different."

2.8 Independent repeated experiments

Our mathematical model of a random experiment is a probability space $(\Omega, \mathcal{F}, P)$. And of the repeated experiment is $(\Omega^2, \mathcal{F}^2, ?)$. The question mark is there because we need to decide the probabilities of events in a repeated experiment. To do this in a simple way we shall consider the case where the *experiments* are independent. That is, the outcome of the first experiment provides no information about the outcome of the second experiment.

Consider a compound event $E_1 \times E_2$, which means that the outcome of the first experiment was in $E_1 \in \mathcal{F}$ and the outcome of the second experiment was in $E_2 \in \mathcal{F}$. The event E_1 in the first experiment is really the event $E_1 \times \Omega$ in the compound experiment. That is, it is the set of all ordered pairs where the first coordinate belongs to E_1. Similarly the event E_2 in the second experiment corresponds to the event $\Omega \times E_2$ in the compound experiment. Now observe that

Add a picture, and some examples.

$$(E_1 \times \Omega) \bigcap (\Omega \times E_2) = E_1 \times E_2.$$

Since the experiments are independent the probability of the intersection $(E_1 \times \Omega)(\Omega \times E_2)$ should be the probability of $(E_1 \times \Omega)$ times the probability of $(\Omega \times E_2)$. But these probabilities are just

$P(E_1)$ and $P(E_2)$ respectively. Thus for independently repeated experiments and "rectangular events,"

$$\text{Prob}(E_1 \times E_2) = P(E_1) \times P(E_2).$$

This is enough to pin down the probability of all the events in the product algebra $\mathcal{F}^2$, and the resulting probability measure is called the product probability, and may be denoted by $P \times P$, or P^2, or by really abusing notation, simply P again.

The point to remember is that independent experiments give rise to products of probabilities.

How do we know when two experiments are independent?

We rely on our knowledge of physics or biology or whatever to tell us that the outcome of one experiment yields no information on the outcome of the other. It's built into our modeling decision. I am no expert, but my understanding is that quantum entanglement implies that certain experiments that our intuition suggests are independent are not really independent.[5] But that is an exceptional case. For coin tossing, die rolling, roulette spinning, etc., independence is probably a good modeling choice.

As an example of experiments that are not independent, consider testing potentially fatal drugs on lab rats, with the same set of rats. If a rat dies in the first experiment, it diminishes the probability he survives through the second.

2.9 ⋆ A digression on infinity

☡ Consider an experiment with a **stopping rule**. For example, consider the experiment, "toss a coin until Heads occurs, then stop." What is the natural sample space, and set of events for this experiment. You might think the simplest sample space for this experiment is the set of sequences of finite length. For a given natural number n, the space of all sequences of length n is the n-fold Cartesian product $\Omega^n = \underbrace{\Omega \times \cdots \times \Omega}_{n \text{ copies}}$, so the grand sample space is

$$\boldsymbol{\Omega} = \bigcup_{n=1}^{\infty} \Omega^n.$$

This sample space is infinite, but at least is a nice infinity—it is countably infinite.

☡☡ The event H_n that the first Head occurs on the n^{th} toss belongs $\mathcal{F}^n$, and so it should be an event in the larger experiment. Now consider the "event"

$$H = (\text{a Head eventually occurs}).$$

The event H is the infinite union $\cup_{n=1}^{\infty} H_n$. Is this union an event as we have defined things? No, it is not. One way to see this is to ask, what is the complement of H? It would be the event that no Head occurs, so we would have to toss forever. But the infinite sequence of all Tails (while admittedly a probability zero occurrence) does not appear in our sample space. Another way to say this is that $\bigcup_{n=1}^{\infty} \mathcal{F}^n$ is not a σ-algebra. So if we want the set of events to include H, we need to do something drastic.

One possibility is never to consider H to be an event. After all, how could we ever "observe" such an event happening? In other words, we could say that ensuring that the set of events is a σ-algebra instead of merely an *algebra* is not worth the trouble. (Many textbooks simply ignore this difficulty, and they are still full of useful results.)

☡☡ On the other hand, we might really care about the probability of the event H. If we want to do that, we may want to agree that the real sample space is actually the set of all infinite sequences of outcomes of the original experiment, that is, the infinite Cartesian product $\boldsymbol{\Omega} = \Omega^{\infty}$.

[5] I asked David Politzer if this was a fair statement, and he gave his blessing.

Hopefully, you remember from Ma 1a that if Ω is countable, then $\bigcup\limits_{n=1}^{\infty} \Omega^n$ is a countable set while Ω^∞ is uncountable if Ω has at least two elements. (Think of binary expansions of real numbers in the unit interval.) This makes $\boldsymbol{\Omega} = \Omega^\infty$ a more difficult object to work with.

I'll have more to say about this when we discuss the Strong and Weak Law of Large Numbers. Each approach has its advantages and disadvantages. I should discuss some of these issues in an appendix for the mathematically inclined, and shall when I can find the time. Fortunately, there are still plenty of interesting things we can say without having to worry about making H an event.

2.10★ Appendix: More on $\boldsymbol{\sigma}$-algebras

This section is to fill in some of the technical details for math nerds. You could probably live a happy and productive life without reading it.

Recall Definition 1.4.5.

1.4.5 Definition *An* ***algebra*** *or* ***field*** *$\mathcal{F}$ of subsets of a set Ω is a set of subsets of$|/$ Ω satisfying:*

1. $\varnothing \in \mathcal{F}$, $\Omega \in \mathcal{F}$.

2. *If $E \in \mathcal{F}$, then $E^{\mathrm{c}} \in \mathcal{F}$.*

3. *If E and F belong to $\mathcal{F}$, then EF and $E \cup F$ belong to $\mathcal{F}$.*

It follows by induction that if $\mathcal{F}$ is an algebra and $E_1, \dots, E_n$ belong to $\mathcal{F}$, then $\bigcap\limits_{i=1}^{n} E_i$ and $\bigcup\limits_{i=1}^{n} E_i$ also belong to $\mathcal{F}$.

A ***$\boldsymbol{\sigma}$-algebra*** *or* ***$\boldsymbol{\sigma}$-field*** *is an algebra of subsets $\mathcal{F}$ that in addition satisfies*

3′. *If $E_1, E_2, \dots$ belong to $\mathcal{F}$, then $\bigcap\limits_{i=1}^{\infty} E_i$ and $\bigcup\limits_{i=1}^{\infty} E_i$ belong to $\mathcal{F}$.*

For example, let $\Omega = \mathbb{Z}$, the set of integers. Some examples of σ-algebras of sets of Ω:

- $\mathcal{F}_1 = \{\mathbb{Z}, \varnothing\}$ is the trivial σ-algebra.

- $\mathcal{F}_2 = \{A : A \subset \mathbb{Z}\}$, the **power set** of $\mathbb{Z}$ is a σ-algebra.

- $\mathcal{F}_3 = \{\mathbb{Z}, \varnothing, E, O\}$, where E is the set of even integers and $O = E^{\mathrm{c}}$ is the set of odd integers, is a σ-algebra.

2.10.1 Definition (Measurable functions) *Given a nonempty set Ω and a σ-field $\mathcal{F}$ of subsets of Ω, an extended real function $f \colon \Omega \to \boldsymbol{R}^\sharp$ is* ***$\boldsymbol{\mathcal{F}}$-measurable*** *if for every interval $A \subset \boldsymbol{R}$, the inverse image (or preimage) of A belongs to $\mathcal{F}$, that is,*

$$\{\omega \in \Omega : f(\omega) \in A\} \in \mathcal{F}.$$

An important and useful fact is this.

2.10.2 Proposition *If $\{\mathcal{F}_i : i \in I\}$, where I is an arbitrary nonempty index set, is a set of σ-algebras of subsets of a set Ω, then*

$$\bigcap_{i \in I} \mathcal{F}_i = \Big\{ A : (\forall i \in I)\ [\, A \in \mathcal{F}_i \,] \Big\}$$

is also a σ-algebra.

The proof is a simple consequence of the definition. For instance, if $A \in \bigcap_{i \in I} \mathcal{F}_i$, then $A \in \mathcal{F}_i$ for each i. But $\mathcal{F}_i$ is a σ-algebra, so $A^{\mathrm{c}} \in \mathcal{F}_i$ (for each i) , so $A^{\mathrm{c}} \in \bigcap_{i \in I} \mathcal{F}_i$. This shows that $\bigcap_{i \in I} \mathcal{F}_i$ is closed under taking complements. And so on.

2.10.3 Corollary *Every (nonempty) set $\mathcal{C}$ of subsets of Ω is included in a unique* smallest *σ-algebra of subsets of Ω, denoted $\sigma(\mathcal{C})$, the* ***σ-algebra generated by $\mathcal{C}$****.*

Proof: Let $I = \{\sigma\text{-algebras } \mathcal{F} : \mathcal{C} \subset \mathcal{F}\}$. Then

$$\sigma(\mathcal{C}) = \bigcap_{\mathcal{F} \in I} \mathcal{F}$$

is a σ-algebra that includes $\mathcal{C}$, and it is clearly included in any other σ-algebra that includes $\mathcal{C}$, so it is smallest. ■

Note that the condition is that $\mathcal{C}$ is a nonempty set of subsets, not a set of nonempty subsets. For instance $\sigma\big(\{\varnothing\}\big) = \{\varnothing, \Omega\}$.

The most important example of a generated σ-algebra is the Borel σ-algebra of subsets of the real numbers.

2.10.4 Definition *The* ***Borel σ-algebra****, denoted $\mathcal{B}$, is the σ-algebra of subsets of $\boldsymbol{R}$ generated by the set of all intervals (closed, open, half-open, degenerate).*

Members of the Borel σ-algebra are called ***Borel sets****.*

A ***Borel function*** *is a function from $\boldsymbol{R}$ to $\boldsymbol{R}$ which is $\mathcal{B}$-measurable, that is, the inverse image of any interval is a Borel set.*

Since the singleton $\{x\}$ is the closed interval $[x]$, every singleton is a Borel set. Consequently, since the Borel σ-algebra is closed under countable unions, it includes every countable set. For instance, the set $\mathbb{Q}$ of rational numbers is a Borel set. Since σ-algebras are closed under complementation, the set $\mathcal{I}$ of irrational numbers is also a Borel set. In fact, without an advanced analysis class, you can't describe a set that is not a Borel set, but they exist.

Recall Definition 2.5.2, which requires that a random variable X have the property that for any interval I,

$$X^{-1}(I) = \{\omega \in \Omega : X(\omega) \in I\} \in \mathcal{F}.$$

In fact, since X^{-1} preserves, unions, intersection, and complements we have the following.

2.10.5 Proposition *If X is a random variable on $(\Omega, \mathcal{F})$, then for every Borel set B,*

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}.$$

There is also a Borel σ-algebra of subsets of $\boldsymbol{R}^{\mathrm{n}}$. A subset $\boldsymbol{I}$ of $\boldsymbol{R}^{\mathrm{n}}$ is an **interval rectangle** if it is of the form

$$\boldsymbol{I} = I_1 \times \cdots \times I_n,$$

where each I_j is an interval,

2.10.6 Definition *The Borel σ-algebra, denoted $\mathcal{B}^n$, is the σ-algebra of subsets of $\boldsymbol{R}^{\mathrm{n}}$ generated by the set of all interval rectangles.*

2.11 ⋆ The $\boldsymbol{\sigma}$-algebra of events generated by random variables

Given a vector $\boldsymbol{X} = (X_1, \dots, X_n)$ of random variables defined on the probability space $(\Omega, \mathcal{F}, P)$ and intervals $I_1, \dots, I_n$,

$$(X_1 \in I_1,\ X_2 \in I_2,\ \dots,\ X_n \in I_n) \text{ is an event.}$$

The smallest σ-algebra that contains all these events, as the intervals I_j range over all intervals, is called the **$\boldsymbol{\sigma}$-algebra of events generated by $\boldsymbol{X}$**, and is denoted

$$\sigma(\boldsymbol{X}) \text{ or } \sigma(X_1, \dots, X_n).$$

2.11.1 Definition *A function $g\colon \Omega \to \boldsymbol{R}$ is **$\boldsymbol{X}$-measurable** if for every interval I, the set $g^{-1}(I)$ belongs to $\sigma(\boldsymbol{X})$.*

The following theorem is beyond the scope of this course, but may be found, for instance, in Aliprantis–Border [1, Theorem 4.41] or M. M. Rao [6, Theorem 1.2.3, p. 4].

2.11.2 Theorem *Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random vector on $(\Omega, \mathcal{F}, P)$ and let $g\colon \Omega \to \boldsymbol{R}$. Then the function g is $\sigma(X_1, \dots, X_n)$-measurable if and only if there exists a Borel function $h\colon \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}$ such that $g = h \circ \boldsymbol{X}$.*

This means there is a one-to-one correspondence between $\boldsymbol{X}$-measurable functions and functions that depend only on $\boldsymbol{X}$. As a corollary we have:

2.11.3 Corollary *The set of $\boldsymbol{X}$-measurable functions is a vector subspace of the space of random variables.*

Bibliography

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

[2] T. M. Cover and J. A. Thomas. 2006. *Elements of information theory*, 2d. ed. Hoboken, New Jersey: Wiley–Interscience.

[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] M. M. Rao. 1981. *Foundations of stochastic analysis*. Mineola, NY: Dover. Reprint of the 1981 Academic Press edition.