J. Zhu

550.420 Introduction to Probability

Spring 2017

1 [She02, Chap. 2] Axioms of Probabiltiy

1.1 [She02, Sect. 2.2] Sample Space and Event

Definition 1 (Sample Space and Event)**.** Sample space Ω is the set containing *all possible* sample points *of interest*. Event $E \subseteq \Omega$ is a set containing *some* sample points. In particular, $\Omega, \varnothing$ are events, and $\Omega \cap \varnothing = \varnothing$, $\Omega \cup \varnothing = \Omega$.

Example 1 (Choosing Marbles)**.** The experimenter gets to pick two balls out of a box that contains one blue ball and one green ball. The sampling procedure is with replacement.

- If the experimenter is interested in the event $E_1 = \{\text{exactly one blue marble is drawn}\}$, then it'd be more convenient to construct the sample space to be $\Omega = \{$ BG , GG, BB $\}$. Whether the blue marble is drawn in the first try or not does NOT matter for E_1.
- On the contrary, if $E_2 = \{\text{a blue marble drawn in the first try}\}$ is of the experimenter's interest, then it'd be wiser to construct the sample space to be $\Omega = \{$ GB, BG , GG, BB $\}$, as the order does matter now.

Remark 1. The construction of the sample space is always of the experimenter's disposal. Build Ω carefully, and use the *combination* and *permutation* in [She02, Chap. 1] cautiously.

Example 2 (Throwing a Dart)**.** $\Omega = \big\{(x, y) : x^2 + y^2 \leq 1\big\}$ is established on the basis of ignoring "missed" darts. Those "missed" darts does not shed light on the accuracy estimation of our interest. That being said, Ω can be built upon the events of our interests ONLY.

Definition 2 (Experiment)**.** Intuitively, an experiment is a testing/observation procedure that can be infinitely repeated and has a well-defined set of possible outcomes (sample space). For simplicity, we assume that the experimenter can have a vision of *all possible* outcomes. That being said, the sample space is known, while the possibility/likelihood of each event still remains unknown. So the observation process becomes necessary, given that *certain* (other than Ω and $\varnothing$) individual outcomes (called *sample point*s) cannot be predicted with certainty.

In frequentist's point of view, the probability (to appear soon) of an event is the *long-run relative frequency* of the occurance of that event in a large number of experiments. The ultimate goal of experiment is to *assess the likelihood of events*.

Proposition 1 (Laws of Set Operations)**.**

Relating the laws of set union and set intersection to the laws of scalar addition and scalar multiplication makes it convenient for memorizing's sake. For scalar operations, we have (1) commutativity: $a_1 + a_2 = a_2 + a_1$, $a_1a_2 = a_2a_1$; (2) associativity: $(a_1 + a_2) + a_3 = a_1 + (a_2 + a_3)$, $(a_1a_2)a_3 = a_1(a_2a_3)$; (3) distributive laws: $(a_1 + a_2)a_3 = a_1a_3 + a_2a_3$ for any $a_1, a_2, a_3 \in \mathbb{R}$.

(i) Commutative Laws. $E_1 \cup E_2 = E_2 \cup E_1$, and $E_1 \cap E_2 = E_2 \cap E_1$.

(ii) Associative Laws. $(E_1 \cup E_2) \cup E_3 = E_1 \cup (E_2 \cup E_3)$, and $(E_1 \cap E_2) \cap E_3 = E_1 \cap (E_2 \cap E_3)$.

(iii) Distributive Laws. $(E_1 \cup E_2) \cap E_3 = (E_1 \cap E_3) \cup (E_2 \cap E_3)$, $(E_1 \cap E_2) \cup E_3 = (E_1 \cup E_3) \cap (E_2 \cup E_3)$.

(iv) DeMorgan's Law. $(A \cup B)^{\mathsf{c}} = A^{\mathsf{c}} \cap B^{\mathsf{c}}$ and $(A \cap B)^{\mathsf{c}} = A^{\mathsf{c}} \cup B^{\mathsf{c}}$

Relatable 1. $(E_1 \cup E_2 \cup E_3)^{\mathsf{c}} = (E_1 \cup E_2)^{\mathsf{c}} \cap E_3^{\mathsf{c}} = E_1^{\mathsf{c}} \cap E_2^{\mathsf{c}} \cap E_3^{\mathsf{c}}$.

By *induction*, DeMorgan's Law can be applied to a *countable* number of events.

The analogy between set operation and scalar operation above can be interpreted further:

- $\mathscr{P}(E_1 \cup E_2) = \mathscr{P}(E_1) + \mathscr{P}(E_2)$ if E_1 and E_2 are *disjoint* (mutually exclusive, $E_1 \cap E_2 = \varnothing$).
- $\mathscr{P}(E_1 \cap E_2) = \mathscr{P}(E_1) \cdot \mathscr{P}(E_2)$ if E_1 and E_2 are *independent*.
  Generally, $\mathscr{P}(E_1 \cap E_2) = \mathscr{P}(E_1) \cdot (E_2 | E_1)$ when they are not necessarily independent—conditional probability will be covered soon.

1.2 [She02, Sects. 2.3–2.4] Axioms of Probability

Definition 3 (Probability Law)**.** Probability $\mathscr{P}(E)$ is tells us how likely the event E is to occur. $\mathscr{P}$ is a *set function* that maps (*measurable*) subsets of Ω to $[0,1]$, which satisfies following axioms:

(i) Normalization. $\mathscr{P}(\Omega) = 1$.

(ii) Nonnegativity. $\forall E \subseteq \Omega$, $0 \le \mathscr{P}(E) \le 1$.

(iii) Countable Additivity. If a *countable* number of events $E_1, E_2, \ldots$ are disjoint, then $\mathscr{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathscr{P}(E_i)$.

Remark 2. $\mathscr{P}(\{\omega\})$ is short for $\mathscr{P}(\omega)$. The probability law always has a *set* as its input. [She02, Sect. 2.6] views probability as a *continuous* set function. Intuitively, probability is a measure of occurrence likelihood (belief).

Further useful properties:

- Complementation. $\mathscr{P}(E) + \mathscr{P}(E^{\mathsf{c}}) = 1$. In particular, $\mathscr{P}(\varnothing) = 0$.
- Monotonicity. $E_1 \subseteq E_2$ implies $\mathscr{P}(E_1) \le \mathscr{P}(E_2)$.

  Proof. $E_2 = E_1 \cup (E_1^{\mathsf{c}} \cap E_2)$ implies $\mathscr{P}(E_2) \ge \mathscr{P}(E_1)$, due to the nonnegativity of $\mathscr{P}(E_1^{\mathsf{c}} \cap E_2)$. □

- Inclusion-Exclusion Theorem. $\mathscr{P}(A_1 \cup A_2) = \mathscr{P}(A_1) + \mathscr{P}(A_2) - \mathscr{P}(A_1 \cap A_2)$.

  Proof. $\mathscr{P}(E_1 \cup E_2) = \mathscr{P}(E_1 \cup (E_1^{\mathsf{c}} \cap E_2)) = \mathscr{P}(E_1) + \mathscr{P}(E_1^{\mathsf{c}} E_2) = \mathscr{P}(E_1) + (\mathscr{P}(E_2) - \mathscr{P}(E_1 E_2))$.
  Remark 3. "Splicing" strategy enables us to make use of countable additivity.

  □

- $\mathscr{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathscr{P}(E_i) - \sum_{i_1 < i_2} \mathscr{P}(E_{i_1} E_{i_2}) + \cdots + (-1)^{r+1} \sum_{i_1 < i_2 < \cdot < i_r} \mathscr{P}(E_{i_1} E_{i_2} \cdots E_{i_r}) + \cdots + (-1)^{n+1} \mathscr{P}(E_1 E_2 \cdots E_n)$

Example 3 (Continuous sample space)**.** $([0,1], \mathcal{B}, \lambda)$.

Example 4 (Discrete sample space (may be infinite))**.** Toss a balanced coin. Build up Ω in your favor.

Example 5. Show that if $\mathscr{P}(E_i) = 1$ for $i = 1, \ldots$, then $\mathscr{P}\left(\bigcap_{i=1}^{n} E_i\right) = 1$.

Proof. By subadditivity, $\mathscr{P}\left(\bigcup_{i=1}^{n} E_i^{\mathsf{c}}\right) \le \sum_{i=1}^{n} \mathscr{P}(E_i^{\mathsf{c}}) = \sum_{i=1}^{\infty} [1 - \mathscr{P}(E_i)] = 0$. Thus, $\mathscr{P}\left(\bigcap_{i=1}^{\infty} E_i\right) = 1 - \mathscr{P}\left(\bigcup_{i=1}^{\infty} E_i^{\mathsf{c}}\right) = 1$. □

2 [She02, Chap. 3] Conditional Probability and Independence

2.1 [She02, Sect. 3.2] Conditional Probabilities

We are calculating probabilities when some partial information concerning the result of an experiment is available to us. So let us make use of it.

Definition 4 (Conditional Probability)**.** Conditional probability of A's occurence given that B has occurred is that $\mathscr{P}\left(A \middle| B\right) = \dfrac{\mathscr{P}\left(A \cap B\right)}{\mathscr{P}\left(B\right)}$ provided that $\mathscr{P}\left(B\right) > 0$.

Remark 4 (Multiplication Rule). $\mathscr{P}\left(A_1 \cap \cdots \cap A_n\right) = \mathscr{P}\left(A_1\right)\mathscr{P}\left(A_2 \middle| A_1\right)\cdots\mathscr{P}\left(A_n \middle| A_1 \cap A_2 \cap \cdots \cap A_{n-1}\right)$. In particular, $\mathscr{P}\left(A \cap B\right) = \mathscr{P}\left(B\right)\mathscr{P}\left(A \middle| B\right) = \mathscr{P}\left(A\right)\mathscr{P}\left(B \middle| A\right)$ on account of the symmetry of two events.

Theorem 1 (Law of Total Probability)**.** If A is an event, and $B_1, \ldots, B_n$ is a *partition* of Ω, then we can compute $\mathscr{P}\left(A\right)$ as $\mathscr{P}\left(A\right) = \sum_{i=1}^{n} \mathscr{P}\left(A \cup B_i\right) = \sum_{i=1}^{n}\left(\mathscr{P}\left(B_i\right)\mathscr{P}\left(A \middle| B_i\right)\right)$.

Proof. A can be partitioned into the union of $A \cap B_i$'s ☐

Example 6 ([She02, Ex. 5m on p. 41] Birthday Problems–Application of First Result in Lecture on 02/20)**.** Suppose that each of N men at a party throws his hat into the center of the room. Afterwards, each man randomly selects a hat. What is the probability that none of the men selects his own hat?

Proof. Let E_i for $i = 1, \ldots, N$ be the event that the ith man selects his own hat. Then the probability that *at least* one of the men selects his own hat is given by *Inclusion-Exclusion Principle* [She02, Prop. 4.4 on p. 31].

$$\begin{aligned}&\mathscr{P}\left(E_1 \cup E_2 \cup \cdots \cup E_n\right)\\ =&\sum_{i=1}^{n}\mathscr{P}\left(E_i\right) - \sum_{i_1<i_2}\mathscr{P}\left(E_{i_1}E_{i_2}\right) + \cdots + (-1)^{r+1}\sum_{i_1<i_2<\cdots<i_r}\mathscr{P}\left(E_{i_1}E_{i_2}\cdots E_{i_r}\right) + \cdots + (-1)^{n+1}\mathscr{P}\left(E_1\cdots E_n\right)\end{aligned}$$

where the summation $\sum_{i_1<i_2<\cdots<i_r}\mathscr{P}\left(E_{i_1}\cdots E_{i_r}\right)$ is taken over all the $\binom{n}{r}$ possible subsets of size r of the set $\{1, \ldots, n\}$.

The event $E_{i_1}\cdots E_{i_n}$ that each of the n men $i_1, \ldots, i_n$ selects his own hat, can occur in any of $(N-n)!$ possible ways. Hence, assuming that all $N!$ possible outcomes are *equally likely*, we see that $\mathscr{P}\left(E_{i_1}\cdots E_{i_n}\right) = \dfrac{(N-n)!}{N!}$. Also, as there are $\binom{N}{n}$ terms in $\sum_{i_1<i_2<\cdots<i_n}\mathscr{P}\left(E_{i_1}\cdots E_{i_n}\right)$, it follows that $\sum_{i_1<i_2<\cdots<i_n}\mathscr{P}\left(E_{i_1}\cdots E_{i_n}\right) = \binom{N}{n}\dfrac{(N-n)!}{N!} = \dfrac{1}{n!}$. Thus,

$$\mathscr{P}\left(\bigcup_{i=1}^{N} E_i\right) = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots + (-1)^{N+1}\frac{1}{N!} \overset{N\to\infty}{\longrightarrow} 1 - e^{-1} \approx 0.632$$

The probability that none of the men selects his own hat is $1 - 1 + \dfrac{1}{2!} - \dfrac{1}{3!} + \cdots + \dfrac{(-1)^N}{N!} \overset{N\to\infty}{\longrightarrow} e^{-1} \approx 0.3679$. ☐

Example 7 ([She02, Ex. 2g on. p. 63]–Still Application of First Result in Lecture on 02/20)**.** It is shown that $\mathscr{P}_N$, the probability that there are no matches when N people randomly select from among their own N hats, is given by $\mathscr{P}_N = 1 - \mathscr{P}\left(\bigcup_{i=1}^{N} E_i\right) = \sum_{i=0}^{N}\dfrac{(-1)^i}{i!}$. What is the probability that exactly k of the N people have matches?

Proof. Letting E denote the event that everyone in this set has a match, and letting G be the event that none of the other $N-k$ people have a match. We have $\mathscr{P}(EG) = \mathscr{P}(E)\,\mathscr{P}(G|\,E)$.

Now, let $F_i, i = 1, \ldots, k$ be the event that the ith member of the set has a match. Then

$$\begin{aligned}\mathscr{P}(E) &= \mathscr{P}(F_1 \cdots F_k)\\ &= \mathscr{P}(F_1)\,\mathscr{P}(F_2|\,F_1)\cdots\mathscr{P}(F_k|\,F_1\cdots F_{k-1})\\ &= \frac{1}{N}\frac{1}{N-1}\cdots\frac{1}{N-k+1}\\ &= \frac{(N-k)!}{N!}\end{aligned}$$

Given that everyone in the set of k has a match, the other $N-k$ people will be randomly choosing among their own $N-k$ hats, so the probability that none of them has a match is equal to the probability of no matches in a problem having $N-k$ people choosing among their own $N-k$ hats. Therefore,

$$\mathscr{P}(G|\,E) = \mathscr{P}_{N-k} = \sum_{i=0}^{N_k} \frac{(-1)^i}{i!}$$

Finally, the probability that a specified set of k people have matches and no one else does is

$$\mathscr{P}(EG) = \frac{(N-k)!}{N!}\mathscr{P}_{N-k}$$

Because there will be exactly k matches if the preceding is true for any of the $\binom{N}{k}$ sets of k individuals, the desired probability is

$$\mathscr{P}(\text{exactly } k \text{ mathces}) = \binom{N}{k}\frac{(N-k)!}{N!}\mathscr{P}_{N-k} = \frac{\mathscr{P}_{N-k}}{k!} \xrightarrow{N\to\infty} \frac{e^{-1}}{k!}$$

□

2.2 [She02, Sect. 3.3] Bayes's Formula

Theorem 2 (Bayes' Rule)**.** For $i = 1, \ldots, n$, $\mathscr{P}(B_i|\,A) = \dfrac{\mathscr{P}(B_i \cap A)}{\mathscr{P}(A)} = \dfrac{\mathscr{P}(B_i)\,\mathscr{P}(A|\,B_i)}{\sum_{i=1}^{n}(\mathscr{P}(B_i)\,\mathscr{P}(A|\,B_i))}$.

Example 8 (Polya's Urn Problem)**.** There are r red marbles and b blue marbles in an urn, draw a marble. Keep track of the color, and then put the marble back, as well as $c > 0$ extra marbles of that color.

- On the first draw, $\mathscr{P}(R_1) = \dfrac{r}{r+b}$ and $\mathscr{P}(B_1) = \dfrac{b}{r+b}$.
- On the second draw, $\mathscr{P}(R_2) = \mathscr{P}(R_1)\,\mathscr{P}(R_2|\,R_1) + \mathscr{P}(B_1)\,\mathscr{P}(R_2|\,B_1) = \dfrac{r+c}{r+c+b}\cdot\dfrac{r}{r+b} + \dfrac{r}{r+b+c}\cdot\dfrac{b}{r+b} = \dfrac{r}{r+b}$, which does not change!

Example 9 (Randomized Response by Warner 1965, for analyzing responses for sensitive questions)**.**

Example 10 (Searching)**.** Dr. Torcaso is searching frantically for his lecture notes, which might be in one of n boxes in his office. For $j = 1, \ldots, n$, let L_j be the event that he fails to discover the lecture notes after desperately (and incompletely) rummaging through box j, and let B_j be the event that the lecture notes are in fact in box j. Suppose that $\mathscr{P}(B_j) > 0$ for each j, and that $\mathscr{P}(L_j|\,B_j) < 1$. Shows that

$$\mathscr{P}(B_j|\,L_j) < \mathscr{P}(B_j) < \mathscr{P}(B_j|\,L_i)\,, i \neq j. \tag{1}$$

Proof. 1. One thing to note is that $\mathscr{P}\left(L_j \mid B_j^{\mathsf{c}}\right) = 1 > \mathscr{P}\left(L_j \mid B_j\right)$, then

$$\begin{aligned}
\mathscr{P}\left(L_j\right) &= \mathscr{P}\left(B_j\right)\mathscr{P}\left(L_j \mid B_j\right) + \mathscr{P}\left(B_j^{\mathsf{c}}\right)\mathscr{P}\left(L_j \mid B_j^{\mathsf{c}}\right) \\
&> \mathscr{P}\left(B_j\right)\mathscr{P}\left(L_j \mid B_j\right) + \mathscr{P}\left(B_j^{\mathsf{c}}\right)\mathscr{P}\left(L_j \mid B_j\right) \\
&= \mathscr{P}\left(L_j \mid B_j\right)
\end{aligned}$$

Therefore, $\dfrac{\mathscr{P}\left(B_j \mid L_j\right)}{\mathscr{P}\left(B_j\right)} = \dfrac{\mathscr{P}\left(L_j \mid B_j\right)}{\mathscr{P}\left(L_j\right)} < 1$.

2. For $i \neq j$, we observe that $\mathscr{P}\left(L_i \mid B_j\right) = 1$. Also, $\mathscr{P}\left(L_i\right) = \mathscr{P}\left(B_i\right)\mathscr{P}\left(L_i \mid B_i\right) + \mathscr{P}\left(B_i^{\mathsf{c}}\right)\mathscr{P}\left(L_i \mid B_i^{\mathsf{c}}\right) < \mathscr{P}\left(B_i\right) + \mathscr{P}\left(B_i^{\mathsf{c}}\right) = 1$. Therefore, $\mathscr{P}\left(B_j \mid L_i\right) = \dfrac{\mathscr{P}\left(B_j\right)\mathscr{P}\left(L_i \mid B_j\right)}{\mathscr{P}\left(L_i\right)} > \mathscr{P}\left(B_j\right)$.

□

Relatable 2. A.2.3. (b) in Homework 2.

Example 11 ([She02, Ex. 3c on p. 67])**.** In answering a question on a multiple-choice test, a student either knows the answer or guesses. let p be the probability that the student knows the answer, and $1-p$ be the probability that the student guesses. Assume that a student who guesses at the answer will be correct with portability $1/m$, where m is the number of multiple-choice alternatives. What is the conditional probability that a student knew the answer to a question given that he or she answered it correctly?

Proof. Let C and K denote, respectively, the events that the student answers the question correctly and the event that he or she actually knows the answer.

$$\begin{aligned}
\mathscr{P}\left(K \mid C\right) &= \frac{\mathscr{P}\left(C \mid K\right)\mathscr{P}\left(K\right)}{\mathscr{P}\left(C \mid K\right)\mathscr{P}\left(K\right) + \mathscr{P}\left(C \mid K^{\mathsf{c}}\right)\mathscr{P}\left(K^{\mathsf{c}}\right)} \\
&= \frac{1 \times p}{1 \times p + \frac{1}{m} \times (1-p)} = \frac{mp}{1 + (m-1)p}.
\end{aligned}$$

□

2.3 [She02, Sect. 3.4] Independent Events

Definition 5 (Independence)**.** Two events $A, B \subseteq \Omega$ are *independent* if $\mathscr{P}\left(A \cap B\right) = \mathscr{P}\left(A\right) \cdot \mathscr{P}\left(B\right)$.

Remark 5. If $\mathscr{P}\left(B\right) > 0$, then $\mathscr{P}\left(A \mid B\right) = \mathscr{P}\left(A\right)$.

Relatable 3. If A and B are independent, then

1) A and B^{c} are independent.

2) A^{c} and B are independent.

3) A^{c} and B^{c} are independent.

Remark 6. Events $A_1, \ldots, A_n$ are independent if *every* subcollection of these n events have the property that the probability of their intersection equals the product of their individual probabilities. For any subset $\{i_1, \ldots, i_k\} \subseteq \{1, 2, \ldots, n\}$, we have $\mathscr{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k}\left(A_{i_j}\right)$. To check that n events are independent, we need to verify $2^n - n - 1$ conditions.

Example 12 ([She02, Ex. 4i on p. 85])**.**

Example 13. Consider families of *exactly* n children, with $n \geq 3$. Let A be the event that a family has chidren of both sexes, and let B be the event that there is at most one girl in the family. Show that the *only* value of n for which the events A and B are independent is $n = 3$, assuming that each child has probability $1/2$ of being a boy.

Proof. $\mathscr{P}(A) = 1 - \mathscr{P}(A^{\mathsf{c}}) = 1 - \left(\frac{1}{2^n} + \frac{1}{2^n}\right) = 1 - \frac{1}{2^{n-1}}$, and $\mathscr{P}(B) = \mathscr{P}(\text{all boys}) + \mathscr{P}(\text{one girl}) = \left(\frac{1}{2}\right)^n + n\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^{n-1} = \frac{n+1}{2^n}$, while $\mathscr{P}(A \cap B) = \mathscr{P}(\text{one girl}) = \frac{n}{2^n}$.

For A and B to be independent, we need $2^{n-1} = n + 1$. A direct substitution shows that this identity does not hold for $n = 2$, but it holes for $n = 3$. We now prove by induction that $2^{n-1} > n + 1$ for $n \geq 4$. This is true for $n = 4$ since $8 = 2^{4-1} > 5 = 4 + 1$. Suppose $k \geq 4$ and $2^{k-1} > k + 1$, then $2^k = 2(2^{k-1}) > 2(k+1) > k + 2$, which completes the induction argument. $\square$

Relatable 4. A.3.10 (b) in Homework 3.

References

[She02] R. Sheldon. *A First Course in Probability*. Pearson Education India, 8 edition, 2002.