ECS315 2017/1 Part II Dr.Prapun

5 Probability Foundations

Constructing the mathematical foundations of probability theory has proven to be a long-lasting process of trial and error. *The approach consisting of defining probabilities as relative frequencies in cases of repeatable experiments leads to an unsatisfactory theory.* The frequency view of probability has a long history that goes back to **Aristotle**. It was not until 1933 that the great Russian mathematician A. N. **Kolmogorov** (1903-1987) laid a satisfactory mathematical foundation of probability theory. He did this by taking a number of axioms as his starting point, as had been done in other fields of mathematics. [21, p 223]

We will try to avoid several technical details[15] [16] in this class. Therefore, the definition given below is not the "complete" definition. Some parts are modified or omitted to make the definition easier to understand.

[15] To study formal definition of probability, we start with the ***probability space*** $(\Omega, \mathcal{A}, P)$. Let Ω be an arbitrary space or set of points ω. Recall (from Definition 1.15) that, viewed probabilistically, a subset of Ω is an ***event*** and an element ω of Ω is a ***sample point***. Each event is a collection of outcomes which are elements of the sample space Ω.

The theory of probability focuses on collections of events, called event ***σ-algebras***, typically denoted by $\mathcal{A}$ (or $\mathcal{F}$), that contain all the events of interest (regarding the random experiment $\mathcal{E}$) to us, and are such that we have knowledge of their likelihood of occurrence. The probability P itself is defined as a number in the range $[0, 1]$ associated with each event in $\mathcal{A}$.

[16] The class 2^{Ω} of all subsets can be too large for us to define probability measures with consistency, across all member of the class. (There is no problem when Ω is countable.)

Definition 5.1. Kolmogorov's Axioms for Probability [12]: A **probability measure**[17] is a real-valued set function[18] that satisfies

P1 ***Nonnegativity***:

$$P(A) \geq 0.$$

P2 ***Unit normalization***:

$$P(\Omega) = 1.$$

P3 ***Countable additivity*** or σ-additivity: For every countable sequence $(A_n)_{n=1}^{\infty}$ of disjoint events,

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

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$$

- The number $P(A)$ is called the ***probability*** of the event A
- The entire sample space Ω is called the ***sure event*** or the ***certain event***.
- If an event A satisfies $P(A) = 1$, we say that A is an ***almost-sure event***.
- A ***support*** of P is any set A for which $P(A) = 1$.

(a.s.)

We write

A occurs a.s.

A occurs w.p. 1

From the three axioms[19], we can derive many more properties of probability measure. These properties are useful for calculating probabilities.

[17] Technically, probability measure is defined on a σ-algebra $\mathcal{A}$ of Ω. The *triple* $(\Omega, \mathcal{A}, P)$ is called a ***probability measure space***, or simply a ***probability space***

[18] A real-valued set function is a function the maps sets to real numbers.

[19] Remark: The axioms do not determine probabilities; the probabilities are assigned based on our knowledge of the system under study. (For example, one approach is to base probability assignments on the simple concept of equally likely outcomes.) The axioms enable us to easily calculate the probabilities of some events from knowledge of the probabilities of other events.

Example 5.2. "Direct" construction of a probability measure: Consider a sample space $\Omega = \{1, 2, 3\}$.

A	$\emptyset$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$P(A)$	0							

5.3. $P(\emptyset) = 0$.

P1 : $P(\emptyset) \geq 0$

Two possibilities: $P(\emptyset) = 0$ ✓ ; $P(\emptyset) > 0$

P3 (disjoint): $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

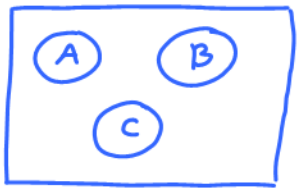
Set all $A_i = \emptyset$

$P(\emptyset) = \sum_{i=1}^{\infty} P(\emptyset)$

contradict

5.4. Finite additivity[20]: If $A_1, \ldots, A_n$ are disjoint events, then

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



The formula is quite intuitive when we use Venn diagram and think of probabilities as areas.

Example: If A_1, A_2, A_3 are disjoint, then $P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$

To see this, start with P3: $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ (don't forget the "disjoint" assumption.)

Set $A_{n+1} = A_{n+2} = A_{n+3} = \ldots = \emptyset$

$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) + 0 + 0 + 0 + \ldots$

Special case when $n = 2$: ***Addition rule*** (Additivity)

$$\text{If } A \cap B = \emptyset, \text{ then } P(A \cup B) = P(A) + P(B). \tag{5}$$

[20] It is not possible to go backwards and use finite additivity to derive countable additivity (P3).

5.5. The probability of a countable event equals the sum of the probabilities of the outcomes in the event.

(a) In particular, if A is countably infinite, e.g. $A = \{a_1, a_2, \ldots\}$, then

$$P(A) = \sum_{n=1}^{\infty} P(\{a_n\}).$$

(b) Similarly, if A is finite, e.g. $A = \{a_1, a_2, \ldots, a_{|A|}\}$, then

$$P(A) = \sum_{n=1}^{|A|} P(\{a_n\}).$$

disjoint

$\{n,u,a\} = \{n\} \cup \{u\} \cup \{a\}$

$A = \{n,u,a\}$ $\quad P(A) = P(\{n,u,a\}) = P(\{n\}) + P(\{u\}) + P(\{a\})$

- This greatly simplifies[21] construction of probability measure.

By defining probability values for singletons, we reduce the # values that we need to specify when we construct P from $\approx 2^n$ values to n values.

Recall

$|2^A| = 2^{|A|}$

$|2^\Omega| = 2^{|\Omega|}$

Remark: Note again that the set A under consideration here is finite or countably infinite. You can not apply the properties above to uncountable set.[22]

---

[21] Recall that a probability measure P is a set function that assigns number (probability) to all set (event) in $\mathcal{A}$. When Ω is countable (finite or countably infinite), we may let $\mathcal{A} = 2^\Omega =$ the power set of the sample space. In other words, in this situation, it is possible to assign probability value to all subsets of Ω.

To define P, it seems that we need to specify a large number of values. Recall that to define a function $g(x)$ you usually specify (in words or as a formula) the value of $g(x)$ at all possible x in the domain of g. The same task must be done here because we have a function that maps sets in $\mathcal{A}$ to real numbers (or, more specifically, the interval $[0, 1]$). It seems that we will need to explicitly specify $P(A)$ for each set A in $\mathcal{A}$. Fortunately, 5.5 implies that we only need to define P for all the singletons (when Ω is countable).

[22] In Section 10, we will start talking about (absolutely) continuous random variables. In such setting, we have $P(\{\alpha\}) = 0$ for any α. However, it is possible to have an uncountable set A with $P(A) > 0$. This does not contradict the properties that we discussed in 5.5. If A is finite or countably infinite, we can still write

$$P(A) = \sum_{\alpha \in A} P(\{\alpha\}) = \sum_{\alpha \in A} 0 = 0.$$

For event A that is uncountable, the properties in 5.5 are not enough to evaluate $P(A)$.

Step: ⓪ Know the probabilities of individual outcomes.

① From the operation(s) that define(s) the event, find the outcomes in the event.

② Add the probabilities of these outcomes

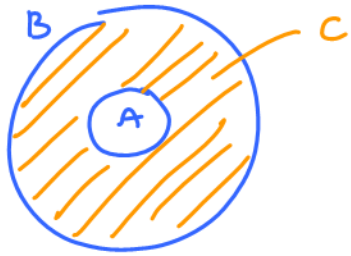
Try these @ home

Example 5.6. A random experiment can result in one of the outcomes $\{a, b, c, d\}$ with probabilities 0.1, 0.3, 0.5, and 0.1, respectively. Let A denote the event $\{a, b\}$, B the event $\{b, c, d\}$, and C the event $\{d\}$.

$P(\{a\}) = 0.1,\ P(\{b\}) = 0.3,\ P(\{c\}) = 0.5,\ P(\{d\}) = 0.1$

- $P(A) = P(\{a, b\}) = P(\{a\}) + P(\{b\}) = 0.1 + 0.3 = 0.4$
- $P(B) =$ P({b,c,d}) = P({b}) + P({c} + P({d})) = 0.3+0.5+0.1 = 0.9
- $P(C) =$ P({d}) = 0.1
- $P(A^c) = P(\{c, d\})$
- $P(A \cap B) = P(\{a, b\} \cap \{b, c, d\}) = P(\{b\}) = 0.3$
- $P(A \cap C) =$ P(∅) = 0

5.7. ***Monotonicity***: If $A \subset B$, then $P(A) \leq P(B)$



$B = A \cup C$ (A ↑ disjoint)

Finite additivity ↓

$P(B) = P(A) + P(C) \Rightarrow P(B) - P(A) = P(C) \geq 0$ ($P(C) \geq 0$ by nonnegativity P1)

$P(B) \geq P(A)$

Example 5.8. Let A be the event to roll a 6 and B the event to roll an even number. Whenever A occurs, B must also occur. However, B can occur without A occurring if you roll 2 or 4.

5.9. If $A \subset B$, then $P(B \setminus A) = P(B) - P(A)$

$B \setminus A = C$

5.10. $P(A) \in [0, 1]$.

P1 ↓ $A \subset \Omega$ ↓ P2

$0 \leq P(A) \leq P(\Omega) = 1$

5.11. $P(A \cap B)$ can not exceed $P(A)$ and $P(B)$. In other words, "the composition of two events is always less probable than (or at most equally probable to) each individual event."

$A \cap B \subset A \Rightarrow P(A \cap B) \leq P(A)$

$A \cap B \subset B \Rightarrow P(A \cap B) \leq P(B)$

$\Leftarrow P(A \cap B) \leq \min\{P(A), P(B)\}$

Example 5.12 (Slides)**.** Experiments by psychologists Kahneman and Tversky.

Example 5.13. Let us consider Mrs. Boudreaux and Mrs. Thibodeaux who are chatting over their fence when the new neighbor walks by. He is a man in his sixties with shabby clothes and a distinct smell of cheap whiskey. Mrs.B, who has seen him before, tells Mrs. T that he is a former Louisiana state senator. Mrs. T finds this very hard to believe. "Yes," says Mrs.B, "he is a former state senator who got into a scandal long ago, had to resign, and started drinking." "Oh," says Mrs. T, "that sounds more likely." "No," says Mrs. B, "I think you mean less likely."

Strictly speaking, Mrs. B is right. Consider the following two statements about the shabby man: "He is a former state senator" and "He is a former state senator who got into a scandal long ago, had to resign, and started drinking." It is tempting to think that the second is more likely because it gives a more exhaustive explanation of the situation at hand. However, this reason is precisely why it is a less likely statement. Note that whenever somebody satisfies the second description, he must also satisfy the first but not vice versa. Thus, the second statement has a lower probability (from Mrs. T's subjective point of view; Mrs. B of course knows who the man is).

This example is a variant of examples presented in the book *Judgment under Uncertainty* [11] by Economics Nobel laureate Daniel Kahneman and co-authors Paul Slovic and Amos Tversky. They show empirically how people often make similar mistakes when they are asked to choose the most probable among a set of statements. It certainly helps to know the rules of probability. A more discomforting aspect is that the more you explain something in detail, the more likely you are to be wrong. If you want to be credible, be vague. [17, p 11–12]

5.14. Complement Rule:

$A \cup A^c = \Omega$

$P(\quad) = P(\Omega)$

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

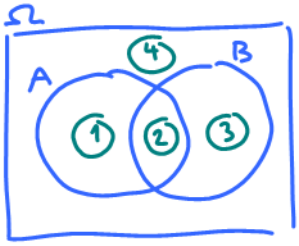
$P(A) + P(A^c) = 1$

- "The probability that something does not occur can be computed as one minus the probability that it does occur."
- Named "probability's Trick Number One" in [10]

5.15. Probability of a union (not necessarily disjoint):

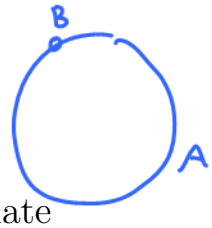
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

(1)+(2) (2) (1)+(2)+(3) (2)+(3)



(1): $P(A \cap B^c)$ (3): $P(B \cap A^c)$

(2): $P(A \cap B)$ (4): $P(A^c \cap B^c)$



- $P(A \cup B) \leq P(A) + P(B)$.
- Approximation: If $P(A) \gg P(B)$ then we may approximate $P(A \cup B)$ by $P(A)$.

1 in 100 1 in 10^9

Example 5.16 (Slides)**.** Combining error probabilities from various sources in DNA testing

Example 5.17. In his bestseller *Innumeracy*, John Allen Paulos tells the story of how he once heard a local weatherman claim that there was a 50% chance of rain on Saturday and a 50% chance of rain on Sunday and thus a 100% chance of rain during the weekend. Clearly absurd, but what is the error?

Answer: Faulty use of the addition rule (5)!

If we let A denote the event that it rains on Saturday and B the event that it rains on Sunday, in order to use $P(A \cup B) = P(A) + P(B)$, we must first confirm that A and B cannot occur at

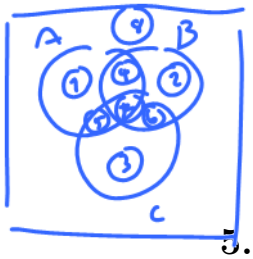
the same time ($P(A \cap B) = 0$). More generally, the formula that is always holds regardless of whether $P(A \cap B) = 0$ is given by 5.15:

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

The event "$A \cap B$" describes the case in which it rains both days. To get the probability of rain over the weekend, we now add 50% and 50%, which gives 100%, but we must then subtract the probability that it rains both days. Whatever this is, it is certainly more than 0 so we end up with something less than 100%, just like common sense tells us that we should.

You may wonder what the weatherman would have said if the chances of rain had been 75% each day. [17, p 12]

5.18. Probability of a union of three events:



$$\begin{aligned} P(A \cup B \cup C) = {} & P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ & + P(A \cap B \cap C) \end{aligned}$$

5.19. Two bounds:

(a) **Subadditivity or Boole's Inequality**: If $A_1, \ldots, A_n$ are events, not necessarily disjoint, then

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

$P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$

(b) **σ-subadditivity** or **countable subadditivity**: If A_1, A_2, ... is a sequence of measurable sets, not necessarily disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$$

- This formula is known as the ***union bound*** in engineering.

$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) + P(A \cap B_4)$

5.20. If a (finite) collection $\{B_1, B_2, \ldots, B_n\}$ is a partition of Ω, then

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

Similarly, if a (countable) collection $\{B_1, B_2, \ldots\}$ is a partition of Ω, then

$$P(A) = \sum_{i=1}^{\infty} P(A \cap B_i)$$

5.21. Connection to classical probability theory: Consider an experiment with **finite** sample space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ in which each outcome ω_i is **equally likely**. Note that $n = |\Omega|$.

$P(\{\omega_i\}) = \beta$ for any ω_i

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

P2 $\Rightarrow$ $1 = P(\Omega) = \sum_{i=1}^{n} P(\{\omega_i\}) = n\beta \Rightarrow \beta = \frac{1}{n} = \frac{1}{|\Omega|}$

We must have

$$P(\{\omega_i\}) = \frac{1}{n}, \quad \forall i.$$

Now, given any ~~event~~ finite[23] event A, we can apply 5.5 to get

$$P(A) = \sum_{\omega \in A} P(\{\omega\}) = \sum_{\omega \in A} \frac{1}{n} = \frac{|A|}{n} = \frac{|A|}{|\Omega|}.$$ ← same formula as Ch. 3.

We can then say that the probability theory we are working on right now is an extension of the classical probability theory. When the conditons/assumptions of classical probability theory are met, then we get back the defining definition of classical classical probability. The extended part gives us ways to deal with situation where assumptions of classical probability theory are not satisfied.

[23] In classical probability, the sample space is finite; therefore, any event is also finite.