Basic Probability Theory[1]

Consider the weather for the upcoming weekend, something relevant for many of us. It is uncertain to us at this point. But suppose that we can list all the possibilities, say sunny, rainy, and cloudy, and suppose that we can imagine saying that there is more chance of one weather than another. Then the uncertainty becomes randomness. It is in the environment of randomness that probability theory is designed for. It is a mathematical language that describes randomness in a precise way.

1 Basic Building Blocks

The basic building blocks of probability theory are *Sample space, Events*, and *Probability Function.*

- Sample Space. Probability theory considers the random object we are concerned about as an experiment. The set S of all possible outcomes of the experiment is called the *sample space.*

  Example: for the weather tomorrow, $S = \{\text{rainy, sunny, cloudy}\}$.

  Example: for inflation rate, $S = R$.

- Event. In probability theory, an event is a collection of outcomes. It is thus a subset of the sample space S.

  Example: $E = \{ \text{ rainy}\}$.

[1] This lecture note is largely adapted from Professor Bruce Hansen's handwritten notes, however, all errors are mine.

Example: $E = (1\%, 2\%)$.

Since an event is a subset of S, one can apply all the set operations, like union ($A \cup B = \{x \in S : x \in A \text{ or } x \in B\}$), intersection ($A \cap B = \{x \in S : x \in A \text{ and } x \in B$), and complement ($A^c = \{x \in S : x \notin A\}$) on events. For example:

- $A \cup B = B \cup A$; $A \cap B = B \cap A$ (commutativity)
- $A \cup (B \cup C) = (A \cup B) \cup C$; $A \cap (B \cap C) = (A \cap B) \cap C$ (associativity)
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$; $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (distributive laws)
- $(A \cup B)^c = A^c \cap B^c$; $(A \cap B)^c = A^c \cup B^c$ (DeMorgan's Laws)

A useful operation for econometrics is as follows and it can be proved using the four basic rules just listed:

$$B = (B \cap A) \cup (B \cap A^c) \tag{1}$$

That is, B can be written as the union of two disjoint sets partitioned by A.

Proof.

$$\begin{aligned}(B \cap A) \cup (B \cap A^c) &= B \cap (A \cup A^c) && \text{(distributive law)} \\ &= B \cap S \\ &= B. \end{aligned} \tag{2}$$

□

2 Probability Function

Probability Function. A probability function P is a function that assigns a value to events and satisfies the **three Axioms of Probability**:

(a) $P(A) \geq 0$ for any $A \subseteq S$.

(b) $P(S) = 1$.

(c) If A and B are disjoint ($A \cap B = \emptyset$), then $P(A \cup B) = P(A) + P(B)$ (finite additivity). When S contains infinite elements, the axiom also requires countable additivity: If $A_1, A_2, \ldots$ are disjoint, then

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

Example: When $S = \{\text{sunny, cloudy,rainy}\}$,

$$P(A) = \begin{cases} 0.2 & A = \{\text{rainy,cloudy}\} \\ 0.8 & A = \{\text{sunny}\} \\ 0 & A = \emptyset \\ 1 & A = S \end{cases} \tag{3}$$

Objective vs. Subjective Probability But what do probabilities mean? There two schools of thoughts here: objective and subjective. In objective probability, probabilities are frequencies. In the weather example, a frequentist imagines a large number of days just like today (same radar reading, wind condition, etc.) and 80% of the days are followed by a sunny weekend. As another example, a frequentist calls a coin fair if it lands head-up 50% of the time when tossed a large number of times. As yet another example, a frequentist says that a college graduate has a 80% probability of earning an income higher than the median income when she randomly draw many college graduates and see that 80% them earn a higher-than-median income.

The frequentist view is very natural when the experiment under consideration can be repeated many times and perhaps have been repeated many times before (like coin tossing, or drawing a random person from a population).

In subjective probability, one assigns probabilities to events based on one's subjective beliefs and information available. Rather than frequency, the probabilities measures the relative likelihood of events in one's belief. This view can be more natural than the frequentist view when the experiment considered cannot be repeated. For example, will Biden be elected president in November? This experiment cannot be repeated and it does not quite help to imagine 1 million parallel universes where the same election will be held. But that does not stop us from assigning a probability to the two possible outcomes, or to feel happy or sad when a renowned prediction agency releases a predicted probability a Biden-win.

It does not have to be an either-or choice between subjective and objective probability. For example, when predicting the probability of a Biden-win, we could use past performance of presidential candidates in a situation similar to Biden's, and we could use polls. Both past performance in similar situlation and polls provide frequentist information. And such information can be used to update one's subjective beliefs.

The Domain of Probability Functions. Note that we were a bit loose when saying that a probability function P assigns a value to *each* event. That sounds like P should assign a value to each and every subset of S, i.e., each set in the power set 2^S of S, but that is incorrect for two reasons:

1. A probability function does not need to assign a value to each and every subset of S. For example, (3) is a perfectly legitimate probability function. The domain of this probability function is $\{\{\text{rainy,cloudy}\}, \{\text{sunny}\}, \emptyset, S\}$.

   In general, a probability function is defined on a collection of subsets, say $\mathcal{B}$, of S that has the following features:

   (a) $\emptyset \in \mathcal{B}$.

   (b) If $A \in \mathcal{B}$, then its complement $A^c \in \mathcal{B}$.

   (c) If $A_1, A_2, \cdots \in \mathcal{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$.

   Or in other words, $\mathcal{B}$ is a σ-field of subsets of S.

   Example: $\mathcal{B}_1 = \{\emptyset, S\}$.

   Example: $\mathcal{B}_2 = \{\{\text{rainy,cloudy}\}, \{\text{sunny}\}, \emptyset, S\}$ where $S = \{\text{sunny, cloudy,rainy}\}$.

   Example: $\mathcal{B}_3 = 2^S := \{A : A \subseteq S\}$.

   All three are σ-fields of subsets of S. Clearly, $\mathcal{B}_1$ is the smallest (coarsest) possible σ-field on S, while $\mathcal{B}_3$ is the largest (finest) possible σ-field of subsets of S. Roughly speaking, probability functions defined on a finer σ-field are more informative than probability functions defined on a coarser one, provided that both are well-defined.

2. It is actually not possible for a probability function to assign a value to each and every subset of S if S is uncountably infinite, like $S = R$.

   When S is uncountably infinite, it is not possible to define a probability function that satisfies all three axioms of probability on $\mathcal{B}_3$. This σ-field is simply too large. It is a perfect weekend past time for some of you geeks out there to Google search for a counter example, but that discussion is technical, non-intuitive, and frankly quite irrelevant for most of econometrics, or all of the econometrics that this course will cover. Thus, we will skip it.

   The commonly used σ-field of subsets of R that probability theory uses is the Borel σ-field, which is the one that we will focus on. The Borel σ-field is the smallest σ-field that contains all open subsets of R. By the definition of σ-field, the Borel σ-field contains all open intervals (a, b), closed intervals $[a, b]$, half-open half-closed intervals $(a, b]$ and $[a, b)$, as well as their finite or countably infinite unions and intersections. That is more than sufficient for all of econometrics applications.

3 Some Properties of Probability Function, Conditional Probability, and Independence of Events

The following properties of probability functions can be easily derived from the three axioms of probability, and are quite useful in econometrics:

1. $P(A) = 1 - P(A^c)$ for all $A \in \mathcal{B}$, where $\mathcal{B}$ is the domain of the probability function P.

   Proof. Note that A and A^c are disjoint, and that $A \cup A^c = S$. Therefore, by the third axiom, we have $P(S) = P(A) + P(A^c)$. Then by the second axiom, $P(A) + P(A^c) = 1$, which proves the claim. □

2. $P(A) \leq 1$.

   Proof. By the first axiom, $P(A^c) \geq 0$. Thus, $P(A) = 1 - P(A^c) \leq 1$. □

3. If $A \subseteq B$, then $P(A) \leq P(B)$.

   Proof. Note that if $A \subseteq B$, then $B = A \cup (B \cap A^c)$. Since A and $B \cap A^c$ are disjoint, we have $P(B) = P(A) + P(B \cap A^c)$ by the third axiom. Also, $P(B \cap A^c) \geq 0$ by the first axiom. Thus, $P(B) \geq P(A)$, proving the claim. □

4. Boole's Inequality: $P(A \cup B) \leq P(A) + P(B)$.

   Proof. Note that $A \cup B = A \cup (B \cap A^c)$. Also, A and $B \cap A^c$ are disjoint. Thus by the third axiom, we have $P(A \cup B) = P(A) + P(B \cap A^c)$. Also, $B \cap A^c \subseteq B$, which implies that $P(B \cap A^c) \leq P(B)$. Therefore, the claim is proved. □

5. Bonferroni's inequality: $P(A \cap B) \geq P(A) + P(B) - 1$.

   Proof. By de Morgan's law, $(A \cap B)^c = (A^c \cup B^c)$. Thus, $P(A \cap B) = 1 - P((A \cap B)^c) = 1 - P(A^c \cup B^c)$ by axiom 3 and property 1 above. Now, the Boole's inequality implies that $P(A^c \cup B^c) \leq P(A^c) + P(B^c)$. Applying property 1 above again, we have $P(A^c) + P(B^c) = 2 - P(A) - P(B)$. Therefore, $P(A \cap B) \geq 1 - (2 - P(A) - P(B)) = P(A) + P(B) - 1$. □

Conditional Probability. Sometimes, we want to know the probability of event B happening given that event A happens. For examle, we might be concerned about the probability of having a rainy day given that the sun will not show up. In this case $B = \{\text{rainy}\}$, and $A = \{\text{cloudy, rainy}\}$. The probability that we are concerned about is the conditional probability of B given A, which is defined as:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

Note that this definition requires $P(A) > 0$.

Example. *Note that the probability function in equation (3) does not provide enough information for us to find $P(B|A)$ when $B = \{\text{rainy}\}$, and $A = \{\text{cloudy, rainy}\}$ because A does not belong to its domain. Let's now consider a different P defined on* *a finer σ-field:*

$$P(E) = \begin{cases} 0.1 & E = \{\textit{rainy}\} \\ 0.1 & E = \{\textit{cloudy}\} \\ 0.8 & E = \{\textit{sunny}\} \\ ... \end{cases} \tag{4}$$

(You can figure out the ... using the three axioms of probability. Hint: there are 5 more members to this finer σ-field.)

Based on this probability function, we can find that $P(A) = P(\{\text{rainy}\}) + P(\{\text{cloudy}\}) = 0.1 + 0.1 = 0.2$, and $P(A \cap B) = P(B) = 0.1$. Therefore $P(B|A) = 0.5$.

Bayes Rule. A particulaly useful formula involving conditional probability is the Bayes rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}.$$

Suppose that $A = \{\text{having disease x}\}$, and $B = \{\text{a screening test for disease x returns postive}\}$. Then the Bayes rule tells you the probabily of the disease given a test result, using the information on the unconditional probability of the disease, the detection rate and the false positive rate of the test.

Bayes rule is the foundation of Bayesian econometrics (an important branch of econometrics which takes the subjective view of probability). Even in frequentist econometrics (a branch of econometrics that takes the objective view of probability), it is useful for calculations and for interpreting results of hypothesis tests.

Proof. Note that $P(B|A)P(A) = P(B \cap A)$, which holds both when $P(A) > 0$ (in which case we use the definition of $P(B|A)$) and when $P(A) = 0$ (in which case both sides are zeros and thus equal). Similarly, $P(B|A^c)P(A^c) = P(B \cap A^c)$. Also note that $B \cap A$ and $B \cap A^c$ are disjoint. Thus by the third axiom of probability, $P(B \cap A) + P(B \cap A^c) = P((B \cap A) \cup (B \cap A^c)) = P(B \cap (A \cup A^c)) = P(B \cap S) = P(B)$. Therefore,

$$\frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \frac{P(B \cap A)}{P(B)} = P(A|B). \tag{5}$$

□

Independence of Events. Sometimes, conditioning on event A does not alter the probability of event B, in which case, we say that A and B are independent events. In math, A and B are independent if $P(A \cap B) = P(A)P(B)$. When $P(A) > 0$, this implies that $P(B|A) = P(B)$.

Assuming certain events are independent allows us to calculate some complicated probabilities via relatively simple probabilities. For example, consider the outcome of rolling two dice. What is the probability of getting a total of 8?

$P(D1 + D2 = 8) = P(D1 = 2, D2 = 6) + P(D1 = 3, D2 = 5) + P(D1 = 4, D2 = 4) + P(D1 = 5, D2 = 3) + P(D1 = 6, D2 = 2) = P(D1 = 2)P(D2 = 6) + P(D1 = 3)P(D2 = 5) + P(D1 = 4)P(D2 = 4) + P(D1 = 5)P(D2 = 3) + P(D1 = 6)P(D2 = 2) = (1/6)(1/6) \times 5 = 5/36$.

Independence of More than Two Events Consider a group of events $A_1, \ldots, A_k$. They are jointly independent if for any subset $J \subseteq \{1, \ldots, k\}$,

$$P(\cap_{j \in J} A_j) = \times_{j \in J} A_j.$$

Notice that this is stronger than pairwise independence (i.e. A_i and A_j are independent for any pair (i, j).

Example. Consider three light bulbs (a,b,c) in a room with closed doors. You are outside the room so do not know the on/off status of the light bulbs. Suppose that the wires are so connected that only four outcomes are possible: all light bulbs are on, only a and b are on, only b and c are on, and only a and c are on, and the four outcomes are equally likely.

Let A be the event that a is on, B be the even tthat b is on, and C be the event that c is on. One can show that A, B, C are pairwise independent but not jointly independent.

4 Problems

1. For two events $A, B \in S$, prove that $A \cup B = (A \cap B) \cup ((A \cap B^c) \cup (B \cap A^c))$.

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

3. Suppose that the unconditional probability of a disease is 0.0025. A screening test for this disease has a detection rate of 0.9, and has a false positive rate of 0.01. Given that the screening test returns positive, what is the conditional probability of having the disease?

4. Suppose that a pair of events A and B are mutually exclusive, i.e., $A \cap B = \emptyset$, and that $P(A) > 0$ and $P(B) > 0$. Prove that A and B are not independent.

5. (Conditional Independence) Sometimes, we may also use the concept of conditional independence. The definition is as follows: let A, B, C be three events with positive probabilities. Then A and B are independent given C if $P(A \cap B|C) = P(A|C)P(B|C)$. Consider the experiment of tossing two dice. Let $A = \{\text{First die is 6}\}$, $B = \{\text{Second die is 6}\}$, and $C = \{\text{Both dice are the same}\}$.

   (a) Show that A and B are independent (unconditionally), but A and B are dependent given C.

   (b) Consider the following experiment: let there be two urns, one with 9 black balls and 1 white balls and the other with 1 black ball and 9 white balls. First randomly (with equal probability) select one urn. Then then take two draws with replacement from the selected urn. Let A and B be drawing a black ball in the first and the second draw, respectively, and let C be the event that urn 1 is selected. Show that A and B are not independent, but are conditionally independent given C.