Economics 361
Probability

Jun Ishii *
Department of Economics
Amherst College

Fall 2023

Overview

Well into the 20^{th} century, scholars debated vigorously on how "probability" should be defined. The first major definition of probability was the following *classical* definition, often credited to Pierre Simon Laplace (early 19^{th} Century).

> *Classical Probability*: If a random experiment can result in N mutually exclusive and equally likely outcomes and if N_A of these outcomes result in the occurrence of the event A, then the probability of A is defined by $P(A) = \frac{N_A}{N}$

This "atomistic" view of probability is also similar to the view held by famed mathematicians Bernoulli, Fermat, and Pascal. Calculations of classical probabilities were essentially exercises in combinatorics.

But this view of probability has several problems. Most notably, it requires the indivisible mutually exclusive outcomes to be "equally likely." Consider the "one die, two dice" problem. A (fair) single die-roll will yield one of 6 mutually exclusive outcomes, each equally likely. But the sum of two (fair) dice rolls has 11 mutually exclusive outcomes which are *not* equally likely.

The idea of using *frequencies* to define probability was attractive. But the requirement of mutually exclusive, equally likely outcomes was not. This led to a modification of the classical definition of probability based on the limiting *relative frequency* of the event.

> *Frequentist Probability*: Let N be the number of times the random experiment is repeated and N_A the number of times event A occurred in the repeated experiments. Then the probability of A is defined by $P(A) = \lim_{N\to\infty} \frac{N_A}{N}$

The classical definition was based on a single random experiment. But the frequentist modification relies on the random experiment being repeated many times (well, infinitely). Probability is the relative frequency of the occurrence event A in an infinite repetition of the random experiment,

*Office: Converse Hall 315 Phone: (413) 542-2901 E-mail: jishii@amherst.edu

in the limit. Therefore, the frequentist definition of probability fails for random experiments that cannot (conceptually) be repeated. This latter point is often overlooked in statistical and econometric studies – sometimes an innocuous omission, other times damning.

This frequentist view of probability is currently the dominant view of probability among practitioners of statistical inference. It is the view that will be maintained throughout most of this course.

One cautionary note: we will be introducing another, separate limit argument later in the course: asymptotic properties of estimators. In the frequentist definition of probability, what is being taken to infinity is the number of repetitions of the random experiment. In discussions about the asymptotic properties of estimators, what is being taken to infinity is the size of the data yielded by the random experiment (the estimation exercise).

An important alternative notion of probability, separate from classical/frequentist, is the notion of *subjective probability*. Subjective probability is the notion closest to the colloquial use of the term probability. Subjective probability reflects the "degree of belief" the observer has on his/her uncertain assessment. Among forms of subjective probability, the most influential is the one utilizing Bayes' Theorem – the Bayesian approach. More details on Bayes' Theorem later.

The Bayesian approach fell out of favor during the latter half of the 20^{th} century due to two main factors. One, scientists were loathe to adopt a practice that could, potentially, differ with each practitioner – not "objective." Two, calculating Bayesian statistics was often very difficult. Recent developments have helped address both complaints, leading some to believe that the Bayesian view may become the dominant view of probability this century.

Probability Theory

The frequentist view of probability can be further formalized using an *axiomatic* approach credited to the Russian mathematician Andrey Kolmogorov. Some preliminary definitions ...

> **DEFINITION**: A **random experiment** is an experiment which satisfies the following conditions
>
> 1. all possible distinct outcomes are known *a priori*
> 2. in any particular trial the outcome is not know *a priori*
> 3. it can be repeated under identical conditions
>
> (from Spanos (1986) Chapter 3)

> **DEFINITION**: The set, S, of all possible outcomes of a particular random experiment is called the **sample space** for the experiment

> **DEFINITION**: An **event** is any collection of possible outcomes of a random experiment, that is, any subset of S (including S itself)
>
> - A **simple event** is an event which cannot be a union of other events
> - A **composite event** is an event which is not a simple event
>
> (from Casella & Berger (1990) Chapter 2 and Amemiya (1994) Chapter 2)

> **DEFINITION**: A collection of subsets of S is called a **Borel field** (or **sigma-algebra**), denoted $\mathcal{B}$, if it satisfies the following properties
>
> 1. $\emptyset \in \mathcal{B}$ (the empty set is contained in $\mathcal{B}$)
> 2. If $A \in \mathcal{B}$ then $A^c \in \mathcal{B}$ ($\mathcal{B}$ is closed under complementation)
> 3. If $A_1, A_2, \cdots \in \mathcal{B}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$ ($\mathcal{B}$ is closed under countable unions)
>
> (from Casella & Berger (1990) Chapter 2)

These preliminary definitions allows us to define, formally, a **probability function**

> **DEFINITION**: Given a sample space S and an associated Borel field $\mathcal{B}$, a **probability function** is a function P with domain $\mathcal{B}$ that satisfies
>
> 1. $P(A) \geq 0$ for all $A \in \mathcal{B}$
> 2. P(S) = 1
> 3. If $A_1, A_2, \cdots \in \mathcal{B}$ are pairwise disjoint, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
>
> (from Casella & Berger (1990) Chapter 2)

The three above conditions are usually referred to as Kolmogorov's Axioms of Probability. The axioms allow us to map every possible outcome (elements in the sample space S), their complements, and their countable unions to some value between 0 and 1.

Furthermore, the axioms ensure the following useful properties of P (where $A, B \in \mathcal{B}$)

- $P(\emptyset) = 0$
- $P(A) \leq 1$
- $P(A^c) = 1 - P(A)$
- $P(B \cap A^c) = P(B) - P(A \cap B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- If $A \subset B$ then $P(A) \leq P(B)$
- $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$ (Boole's Inequality)
- $\sum_{i=1}^{\infty} P(A \cap C_i) = P(A)$ for any partition $C_1, C_2, \ldots$

Aside: Some Basic Set Theory

Set notation:

- $A \subset B$: "A is contained by B and is a subset of B." If $x \in A$ then $x \in B$
- $A = B$: "A is equal to B." $A \subset B$ and $B \subset A$
- $A \cup B$: "Union of A and B." $A \cup B = \{x : x \in A \text{ or } x \in B\}$
- $A \cap B$: "Intersection of A and B." $A \cap B = \{x : x \in A \text{ and } x \in B\}$
- A^c: "Complement of A." $A^c = \{x : x \notin A\}$

Properties of set operations:

- Commutativity: $A \cup B = B \cup A$ and $A \cap B = B \cap A$
- Associativity: $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$
- Distributive Laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- DeMorgan's Laws: $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$

Some additional definitions

- A and B are **disjoint** (or **mutually exclusive**) if $A \cap B = \emptyset$
- $A_1, A_2, \ldots$ are **pairwise disjoint** (or **mutually exclusive**) if $A_i \cap A_j = \emptyset$ for all $i \neq j$
- If $A_1, A_2, \ldots$ are pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = S$ then the collection $A_1, A_2, \ldots$ forms a **partition** of S

Also note that $A \cup \emptyset = A$

Example: Birthday Problem

Problem: Consider a room with twenty people, all strangers to each other. What is the probability that no two persons share the same birthday? (ignore leap years)

Here, think of the random experiment as that of twenty births (rather than 20 random experiments, each a single birth).

A Classical non-Axiomatic Solution

The classical probability scholar would say that there are 365^{20} equally likely, mutually exclusive outcomes to this experiment. So $N = 365^{20}$. (This is the combinatorics problem of assignment with replacement)

Of the 365^{20} outcomes, only $365 \times 364 \times 363 \times \cdots \times 347 \times 346 = \frac{365!}{345!}$ of those outcomes are associated with the event of "no two persons share the same birthday." So $N_A = \frac{365!}{345!}$. (This is the combinatorics problem of assignment without replacement)

Therefore $P(\text{"no two persons share the same birthday"}) = \frac{N_A}{N} = \frac{365!/345!}{365^{20}} \approx 0.59$.

A Classical Axiomatic Solution

Suppose 19 of the birthdays are different (e.g. January 1-19). How many outcomes have those 19 birthdays *and* a different 20^{th}? $365 - 19 = 346$. This is true for any set of 19 different birthdays (under the classical assumption that each unique set of 19 birthdays is equally likely).

There are $365 \times 364 \times \cdots \times 348 \times 347 = \frac{365!}{346!}$ possible sets of 19 different birthdays. So we can partition the event of "no two persons share the same birthday" into $\frac{365!}{346!}$ events. Denote each such partition as C_i and the event "no two persons share the same birthday" as A.

From the axioms, $P(A) = \sum_{i=1}^{365!/346!} P(A \cap C_i)$

From above, $P(A \cap C_i) = \frac{346}{365^{20}}$

Therefore $P(A) = \sum_{i}^{365!/346!} \frac{346}{365^{20}} = \left(\frac{365!}{346!}\right) \times \left(\frac{346}{365^{20}}\right) = \frac{365!/345!}{365^{20}} \approx 0.59$.

A Frequentist Solution?

The above classical probability solutions require each possible disjoint set of 20 birthdays to be equally likely. This may not be an innocuous assumption – some birthdays may be more likely than others. The frequentist probability scholar can relax this assumption, as probability is defined as the limiting relative frequency of the event. But calculating the actual frequentist probability of the event requires other assumptions as there are no general combinatorics rules for $\lim_{N\to\infty} \frac{N_A}{N}$.

Conditional Probability

An important category of event is of the form "event A occurs given that event B will occur." This is different from the joint event "event A *and* B occurs." We denote the former as $A|B$ and the latter as $A \cap B$.

The complement of $A \cap B$, $(A \cap B)^c$, includes A^c and B^c. But the complement of $A|B$, $(A|B)^c$ does *not* include B^c. It only includes $(A^c \cap B)$. This is because "$|B$" (read "conditioning on B") limits the sample space only to events where B occurs.

The conditional probability function of A given B is defined as

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)} \quad \text{for } P(B) > 0$$

which is the ratio of the probability function of the joint event $A \cap B$ and the probability function of the conditioning event B (which is a possible).

The "appropriateness" of this definition is perhaps clearest under the classical definition of probability. $P(A \cap B) = \frac{N_{A \cap B}}{N}$ and $P(B) = \frac{N_B}{N}$. So $P(A|B) = \frac{N_{A \cap B}}{N_B}$, the ratio of the number of outcomes where both A and B occur and the number of outcomes where B occurs. This is the probability of A for a sample space where B always occurs: what "conditioning" does is limit the sample space.

The following is another way to see the "appropriateness" of the definition

> Consider the possible event B. By the axioms of probability
>
> $$\begin{aligned} P(B) &= P(B \cap A) + P(B \cap A^c) && \text{Note: } (A, A^c) \text{ is a proper partition} \\ \frac{P(B)}{P(B)} &= \frac{P(A \cap B)}{P(B)} + \frac{P(A^c \cap B)}{P(B)} && \text{Note: } P(B) > 0, \text{ Commutativity} \\ 1 &= P(A|B) + P(A^c|B) && \text{by definition of conditional probability} \end{aligned}$$

Once the sample space is limited to events where B occurs ("conditioning on B"), the sum of the conditional probability of event A and of all other events (A^c) is 1, as desired.

More formally, the definition of conditional probability function can be derived from its own set of axioms; see Amemiya (1994) Ch.2.4.1 for more details. The properties for a probability function listed earlier also apply for a conditional probability function.

Additionally, the conditional probability function may be concatenated

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

Note: $B, C = B \cap C$. To see that the above holds, let $H = B \cap C$ and consider $P(A|H)$.

Statistical Independence

Relationship among events play an important role in statistical inference. When one event, say A, necessarily entails the other, say B, then A is a subset of B: $A \subset B$. $P(A) \leq P(B)$, $P(A \cap B) = P(A)$, and $P(B|A) = 1$. Alternatively, when one event, A, necessarily precludes the other, B, $P(A \cap B) = 0$ and $P(A|B) = P(B|A) = 0$. More often, two events A and B neither entail nor preclude the other: $P(B|A) \in (0, 1)$.

The special case where the relationship between A and B is such that $P(B|A) = P(B)$. Conditioning on B does not change the probability of A; the probability of A is the same whether we use the full sample space or the sample space limited to events where B occurs. In this situation, A and B are said to be *statistically independent* of each other.

> **DEFINITION**: Events A and B are said to be (pairwise) **statistically independent** if $P(A|B) = P(A)$

Note: $P(A|B) = P(A)$ implies (i) $P(A \cap B) = P(A)P(B)$ and (ii) $P(B|A) = P(B)$.
The definition of statistical independence can be expanded to the comparison of $n > 2$ events

> **DEFINITION**: Events $A_1 \dots A_n$ are said to be **mutually independent** if any proper subset of the events are mutually independent and $P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2) \cdots P(A_n)$

Note that pair-wise independence does not imply mutual independence. Even if all possible pairs of the $n > 2$ events are pairwise independent, the n events may not be mutually independent.

Consider the following counter-example (from Amemiya (1994) Chapter 2.4). Let the random experiment be two fair coin flips (here, it does not matter whether we follow a classical or frequentist view). Let A be the event that H appears on the first coin flip, B the event that H appears in the second coin flip, and C the event that either both flips yield H or both flips yield T. Using combinatorics

- $P(A) = P(B) = P(C) = \frac{2}{4} = \frac{1}{2}$
- $P(A \cap B) = P(A \cap C) = P(B \cap C) = \frac{1}{4}$
- $P(A|B) = P(B|A) = P(A|C) = P(B|C) = P(C|A) = P(B|A) = \frac{1/4}{1/2} = \frac{1}{2}$
- $P(A|B, C) = P(B|A, C) = P(C|A, B) = 1$

So $\{A, B, C\}$ are pairwise independent but not mutually independent. While no individual event entails the other, any two of the events jointly entails the other.

Note: one can show that mutual independence does imply pairwise independence

Example: Birthday Problem, Revisited

Without loss of generality, let the 20 people in the room be indexed from 1 to 20. Let A_i be the event where the i^{th} person has a different birthday than the previous $i-1$ people. Let A denote the event of "no two persons share the same birthday" in the twenty person room. Then

$$\begin{aligned} A &= A_2 \cap A_3 \cap \cdots \cap A_{19} \cap A_{20} \\ P(A) &= P(A_2 \cap A_3 \cap \cdots \cap A_{19} \cap A_{20}) \end{aligned}$$

Using the definition of conditional probability

$$\begin{aligned} P(A) &= P(A_3 \cap \cdots \cap A_{20} | A_2) \times P(A_2) \\ &= P(A_4 \cap \cdots \cap A_{20}) | A_2, A_3) \times P(A_3|A_2) \times P(A_2) \\ &= P(A_{20}|A_2, \ldots, A_{19}) \times \cdots \times P(A_4|A_2, A_3) \times P(A_3|A_2) \times P(A_2) \end{aligned}$$

Under the classical view of probability

$$\begin{aligned} P(A_2) &= \frac{365^{19} \times 364}{365^{20}} = \frac{364}{365} \\ P(A_3|A_2) &= \frac{365^{18} \times 364 \times 363}{365^{19} \times 364} = \frac{363}{365} \\ &\vdots \\ P(A_{20}|A_2, \ldots, A_{19}) &= \frac{346}{365} \end{aligned}$$

Therefore

$$P(A) = \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{346}{365} = \frac{365!/345!}{365^{20}} \approx 0.59$$

Statistical Independence and a Frequentist Solution

Let $A_{i,j}$ denote the event that person i has a birthday j. e.g. $A_{1,1}$ denotes person 1 being born January 1st (day 1). For the birthday problem, the classical probability view of equally likely outcomes essentially assumes

1. $P(A_{i,j}) = P(A_{k,l})$ for $i, j = 1 \ldots 20$ and $k, l = 1 \ldots 365$
2. $\{A_{i,j}\}_{i=1,20 \; j=1,365}$ are mutually independent

Under the frequentist view, the first assumption can be relaxed and $P(A)$ still be calculated if $P(A_{i,j})$ is known for $i = 1$ to 20 and $j = 1$ to 365. The above conditional probability approach can be used as the relevant conditional probabilities can be calculated from $\{P(A_{i,j})\}_{i=1 \text{ to } 20 \; j=1 \text{ to } 365}$.

HINT: A_2 can be partitioned into $(A_2 \cap A_{1,1}) \ldots (A_2 \cap A_{1,365})$
and $A_2 \cap A_{1,j}$ into $(A_{1,j} \cup A_{2,1}) \ldots (A_{1,j} \cup A_{2,j-1})(A_{1,j} \cup A_{2,j+1}) \ldots (A_{1,j} \cup A_{2,365})$
and so forth ...

(No need to solve explicitly – really tedious problem!)

References

Amemiya, Takeshi. *Introduction to Statistics and Econometrics.* Harvard University Press, 1994.

Casella, George & Roger Berger. *Statistical Inference.* Duxbury Press, 1990.

Spanos, Aris. *Statistical Foundations of Econometric Modelling.* Cambridge University Press, 1986.