Wasserman: All of Statistics

Kenny Workman

March 22, 2025

1 Probability

1.1 Basics

Definition 1.1. The **sample space** Ω is the set of outcomes from an experiment. Each point is denoted ω and subsets, eg. $A \subset \Omega$ are called **events**.

Definition 1.2 (Axioms of Probability)**.** A function $\mathbb{P} : \Omega \to \mathbb{R}$ that assigns a real number to each event $A \subset \Omega$ is called a **probability function** or **probability measure** if it satsifies these three axioms:

1. **Non-negativity**. $\mathbb{P}(A) \geq 0$ for every event A
2. **Normalization**. $\mathbb{P}(\Omega) = 1$.
3. **Additivity**. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ if $A \cap B = \emptyset$.

It is incredible, and not obvious, that much of probability is built up from these only these three axioms

Example 1.3. It's actually tricky to show $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ with these three facts:

$$\begin{aligned}
\mathbb{P}(A \cup B) &= \mathbb{P}(AB^c \cap AB \cap A^cB) \\
&= \mathbb{P}(AB^c) + \mathbb{P}(AB) + \mathbb{P}(A^cB) \\
&= \mathbb{P}(AB^c) + \mathbb{P}(AB) + \mathbb{P}(A^cB) + P(AB) - P(AB) \\
&= \mathbb{P}(AB^c \cup AB) + \mathbb{P}(A^cB \cup AB) - P(AB) \\
&= \mathbb{P}(A) + \mathbb{P}(B) - P(AB)
\end{aligned}$$

Another simple idea is that events that are identical at the limit should have identical probabilities.

Theorem 1.4 (Continuity of Events)**.** *If* $A_n \to A$ *then* $\mathbb{P}(A_n) \to \mathbb{P}(A)$.

Proof. Let A_n be monotone increasing: $A_1 \subset A_2 \subset \ldots$. Let $A = \lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i$.

Construct disjoint sets B_i from each A_i where $B_1 = A_1$ and $B_n = \{\omega \in \Omega : \omega \in A_n, \omega \notin \bigcup_{i=1}^{i-1} A_i\}$. It will be shown that (1) each pair of B_i are disjoint, (2) $\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i$ and (3) $A = \bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$ (Exercise 1.1).

From Axiom 3: $\mathbb{P}(A_n) = \mathbb{P}(\bigcup\limits_{i=1}^{n} A_i) = \mathbb{P}(\bigcup\limits_{i=1}^{n} B_i) = \sum\limits_{i=1}^{n} \mathbb{P}(B_i)$.

Then $\lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \sum^{n} \mathbb{P}(B_n) = \sum^{\infty} \mathbb{P}(B_n) = \mathbb{P}(\bigcup^{\infty} B_n) = \mathbb{P}(A)$

□

Definition 1.5 (Conditional Probability)**.** If $\mathbb{P}(B) > 0$, then the probability of A given B is

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}.$$

Exercise 1.1. Fill in the details for Theorem 1.2 and extend to the case where A_n is monotone decreasing.

Proof. For any pair B_{n+1} and B_n, because $B_n \subset A_n$ and $B_{n+1} \cap A_n = \emptyset$, it follows that $B_{n+1} \cap B_n = \emptyset$.
Let $\bigcup_{i=1}^n B_i = \bigcup_{i=1}^n A_i$. Then $\bigcup_{i=1}^{n+1} B_i = (A_{n+1} \setminus \bigcup_{i=1}^n A_i) \bigcup (\bigcup_{i=1}^n A_i) = \bigcup_{i=1}^{n+1} A_i$.
For the monotone decreasing case, let A_n be a sequence where $A_1 \supset A_2 \supset A_3 \ldots$.
Observe $A_1^c \subset A_2^c \ldots$ and $\lim_{n\to\infty} A_n = \Omega \setminus \bigcup^\infty A_i^c$. Construct disjoint B_n^c from A^c in the same way.
Then $\lim_{n\to\infty} \mathbb{P}(A_n) = 1 - \sum^\infty \mathbb{P}(B_i^c) = 1 - \mathbb{P}(A^c) = \mathbb{P}(A)$ □

Exercise 1.3. Let Ω be a sample space and $A_1, A_2, \ldots$ be events. Define $B_n = \cup_{i=n}^\infty A_i$ and $C_n = \cap_{i=n}^\infty A_i$.

(a) Show $B_1 \supset B_2 \supset B_3 \ldots$ and $C_1 \subset C_2 \subset C_3 \ldots$

(b) Show $\omega \in \cap_{n=1}^\infty B_n$ iff ω is in an infinite number of the events

(c) Show $\omega \in \cup n = 1^\infty C_n$ iff ω belongs to all of the events, except possibly a finite number of those events.

Proof. (a) Certainly $\cup_{i=1}^\infty A_i \supset \cup_{i=2}^\infty A_i \ldots$ and $\cap_{i=1}^\infty A_i \subset \cap_{i=2}^\infty \ldots$.

(b) Forward. Assume $\omega \in \cap_{n=1}^\infty B_n$. If ω does not belong to an infinite number of events A_i, there exists some index j past which $\omega \notin B_j$. Then certainly $\omega \notin \cap_{n=1}^\infty B_n$. Reverse. ω belonging to infinite events means there cannot exist such a j described previously so $\omega \in B_n$ for all n. Indeed $\omega \in \cap_{n=1}^\infty B_n$

(c) Forward. Assume $\omega \in \cup_{n=1}^\infty C_n$. Then $\omega \in C_j = \cap_{i=j}^\infty A_i$ for some j. This is another way of saying ω is in every single event except for perhaps a finite number in $A_{i<j}$. Reverse. Let j be the index of the largest event that ω is not in. Then $\omega \in C_{n>j}$ and certainly $\omega \in \cup^\infty C_n$.

□

Note. The key idea above is this notion of "infinitely often" (i.o.) and "all but finitely often" (eventually) which are two distinct structures of infinite occurence in sequences. Consider an ω that exists in every other event (eg. just the odd indices) for infinite events and revisit its inclusion in $\cap^\infty B_i$ and $\cup^\infty C_i$.

Note. $\lim \cap \cup A_n$ is also referred to as the limit infimum of A_n. Similarly, $\lim \cup \cap A_n$ is referred to as the limit supremum of A_n.

Exercise 1.7. Let $\mathbb{P}(\bigcup^n A_i) \leq \sum^n \mathbb{P}(A_i)$. Then $\mathbb{P}(A_{n+1} \cup (\bigcup^n A_i)) \leq \mathbb{P}(A_{n+1}) + (\sum^n \mathbb{P}(A_i)) - \mathbb{P}(A_{n+1} \cap (\bigcup^n A_i)) \leq \sum^{n+1} \mathbb{P}(A_i)$

Note. Expand a bit on the Boole inequality.

Exercise 1.9. For fixed B s.t. $\mathbb{P}(B) > 0$, show $\mathbb{P}(. \mid B)$ satisfies the three axioms of probability.

Proof.
- **Non-negativity**. If $\mathbb{P}(B) > 0$ and $\mathbb{P}(AB) > 0$ for any $A \subset \Omega$, certainly $\frac{\mathbb{P}(AB)}{\mathbb{P}(B)} > 0$.
- **Normalization**. $\frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$
- **Additivity**. Let $AB \cap CB = \emptyset$, then $\mathbb{P}(AB \cap CB) = \mathbb{P}(AB) + \mathbb{P}(CB)$. Indeed $\frac{\mathbb{P}(AB \cap CB)}{B} = \frac{\mathbb{P}(AB)}{B} + \frac{\mathbb{P}(CB)}{B}$

□

Exercise 1.11. Suppose A and B are independent events. Show that A^c and B^c are also independent.

Proof. We are given $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$. Then $\mathbb{P}(A^c)\mathbb{P}(B^c) = (1 - \mathbb{P}(A))(1 - \mathbb{P}(B)) = 1 - \mathbb{P}(A) - \mathbb{P}(B) + \mathbb{P}(A)\mathbb{P}(B) = 1 - \mathbb{P}(A \cup B) = \mathbb{P}(A^c B^c)$. The second to last equality uses independence of $P(AB)$. The last equality uses the property of set complements $P(A \cup B) = P(A^c \cap B^c)$. □

Exercise 1.13. Suppose a fair coin is tossed repeatedly until heads and tails is each encoutered exactly once. Describe Ω and compute the probablity exactly three tosses are needed.

Proof.
- The sample space is the set of binary strings with exactly one 0 and 1. For strings of length greater than 2, these are repeated strings of 0 or 1 capped with a 1 or 0 respectively.

- By independence, each n-string has an identical probability $\frac{1}{2}^n$. There are two such 3-strings: 001 and 110. Using additivity, $\mathbb{P}(3 \text{ tosses}) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$

$\square$

Exercise 1.15. The probability a child has blue eyes is $\frac{1}{4}$. Assume independence between children. Consider a family with 3 children.

- If it is known that at least one of the children have blue eyes, what is the probablity that at least two of the children have blue eyes?
- If it is know that the youngest child has blue eyes, what is the probability that at least two of the children have blue eyes?

Proof.
- Straightforward conditional probability. Let A be the event where at least one child has blue eyes and B be the event where at least two children have blue eyes. Consider first, $\mathbb{P}(A) = 1 - \mathbb{P}(\text{no child has blue eyes}) = 1 - \frac{27}{64} = \frac{37}{64}$. Compute $\mathbb{P}(A \cap B)$ by enumerating events 101, 111, 110 and using additivity: $2 \cdot \frac{1}{4}^2 \cdot \frac{3}{4} + \frac{1}{4}^3 = \frac{10}{64}$. Then $\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{10}{64} \cdot \frac{64}{37} = \frac{10}{37}$
- Similar procedure. Let A be the event where the youngest child has blue eyes and B be as before. Using independence, $\mathbb{P}(A) = \frac{1}{4}$. (To see this rigorously, enumerate the sample space and see $\mathbb{P}(\Omega \mid \text{first child blue}) = 1$). Now $\mathbb{P}(B \cap A)$ describe events 110, 101, 111 only. $\frac{7}{64} \cdot \frac{4}{1} = \frac{7}{16}$.

$\square$

Exercise 1.17. Show $\mathbb{P}(ABC) = \mathbb{P}(A \mid BC)\mathbb{P}(B \mid C)\mathbb{P}(C)$

Proof. By straightforward application of the definition of conditional probability: $\dfrac{\mathbb{P}(ABC)}{\mathbb{P}(BC)} \dfrac{\mathbb{P}(BC)}{\mathbb{P}(C)} \mathbb{P}(C) = \mathbb{P}(ABC)$ $\square$

Theorem 1.6 (Total Probability)**.** *If $A_1 \cdots A_k$ partition Ω, $\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)$*

Exercise 1.19. Suppose 50% of computer users are Windows. 30% are Mac. 20% are Linux. Suppose 65% of Mac users, 82% of Windows users and 50% of Linux users get a virus. We select a person at random and learn they have the virus. What is the probability they are a Windows user?

Proof. Let each $\omega \in \Omega$ be a distinct user. Then $W, M, L \subset \Omega$ are the users with Windows, Mac + Linux machines. $V, N \subset \Omega$ are the users with and without viruses.

We want $\mathbb{P}(W \mid V) = \dfrac{\mathbb{P}(V \mid W)\mathbb{P}(W)}{\mathbb{P}(V)}$. Compute $\mathbb{P}(V) = \sum_{X=\{W,M,L\}} \mathbb{P}(V \mid X)\mathbb{P}(X) = 0.705$. Then $\mathbb{P}(W \mid V) = \dfrac{0.82 \cdot 0.50}{0.705} = 0.581$. $\square$

Exercise 1.20. A box contains 5 coins, each with a different probability of heads: $0, 0.25, 0.5, 0.75, 1$. Let C_i be the event with coin i and H_i be the event that heads is recovered on toss i.Suppose you select a coin at random and flip it.

- What is the posterior probability $\mathbb{P}(C_i \mid H_1)$ for each coin?
- What is $\mathbb{P}(H_2 \mid H_1)$?
- Let B_i be the event that the first heads is recovered on flip i. What is $\mathbb{P}(C_i \mid B_i)$ for each coin?

Proof.
- $\mathbb{P}(H_1) = \frac{1}{2}$. For each coin, $\mathbb{P}(C_i \mid H) = \dfrac{\mathbb{P}(H \mid C_i)\mathbb{P}(C_i)}{\mathbb{P}(H)}$. $\mathbb{P}(H)$ can be worked out using total probability: $\sum_i \mathbb{P}(H|C_i)\mathbb{P}(C_i) = \frac{1}{2}$. Then eg. the posterior $\mathbb{P}(C_4 \mid H) = \frac{3}{4} \cdot \frac{1}{5} \cdot \frac{2}{1} = \frac{3}{10}$.
- Note that both tosses are conditionally independent: $\mathbb{P}(H_2 H_1 \mid C_i) = \mathbb{P}(H_2 \mid C_i)\mathbb{P}(H_1 \mid C_i)$. $\mathbb{P}(H_2 \mid H_1) = \dfrac{\mathbb{P}(H_2 H_1)}{\mathbb{P}(H_1)} = \dfrac{\sum_i \mathbb{P}(H_2 H_1 \mid C_i)\mathbb{P}(C_i)}{\sum_i \mathbb{P}(H_1 \mid C_i)\mathbb{P}(C_i)}$. Because $\mathbb{P}(C_i)$ is uniform, we can simply to $\dfrac{\sum_i \mathbb{P}(H_2 H_1 \mid C_i)}{\sum_i \mathbb{P}(H_1 \mid C_i)}$. The result is $\dfrac{\sum_i p_i^2}{\sum_i p_i}$.

- Similar idea to (a).

$\square$

Note. Important to see that independent events are not conditionally independent in general. Try to construct an example.

1.2 Random Variables

Definition 1.7 (random variable)**.**

Definition 1.8 (cumulative distribution function)**.**

The CDF contains "all the information" in a random variable. This is articulated by the following theorem:

Theorem 1.9. *For random variables X and Y with CDFs F and G, if $F(x) = G(X) \forall x \in [0,1]$, then $X = Y$ ($\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for each $A \subset \mathbb{R}$).*

And the behavior of the CDF, including "all of its information" is uniquely determined by just three properties:

Theorem 1.10. *A function $F : \mathbb{R} \to [0,1]$ is a CDF iff it satisfies three properties:*

- *Non-decreasing.* $x_2 > x_1 \implies F(x_2) \geq F(x_1)$
- *Normalization.* $\lim_{y\to 0} F(y) = 0$ *and* $\lim_{y\to 1} F(y) = 1$
- *Right-continuous. For any $x \in \mathbb{R}$, $F(x) = F^+(x)$ where $F^+(x) = \lim_{y\to x, y>x} F(y)$*

Proof. Starting with (iii) from the text, let $A = (-\infty, x]$ and $y_1, y_2, \ldots$ be a sequence where $y_1 < y_2 \ldots$ and $\lim_i y_i = x$. By the definition of the CDF, $F(y_i) = \mathbb{P}(A_i)$ and $F(x) = \mathbb{P}(A)$, where $\lim_i F(y_i)$ is equivalent to $\lim_{y\to x, y>x} F(y)$. Observe $\cap_i A_i = A$ so $\mathbb{P}(A) = \mathbb{P}(\cap_i A_i) = \lim_i \mathbb{P}(A_i) = \lim_i F(y_i) = F(x)$ as desired.

To see (ii), $\lim_{y\to -\infty} F(y) = 0$, define a sequence $y_1, y_2 \cdots$ where $y_1 > y_2 \cdots$ as before and $y_1 = y$. Let $A_i = (\infty, y_i]$. Then $\cap_i A_i = \emptyset$ and $\mathbb{P}(\cap_i A_i) = \mathbb{P}(\emptyset) = 0$. Indeed $\lim_{y\to -\infty} F(y) = \lim_i \mathbb{P}(A_i) = \mathbb{P}(\cap_i A_i) = 0$. A similar argument shows the limit to the other direction.

For (iii), if $x_2 > x_1$ then $P((-\infty, x_2]) \geq P((-\infty, x_1])$ and $F(x_2) \geq F(x_1)$. $\square$

The interesting direction is the reverse: a function satisfying these properties uniquely determines a probability function. It is difficult to show in general. A concrete example is the Cantor function (Devil's staircase) which satisfies non-decreasing, normality and right-continuous properties but from which is difficult to derive a measure that satisfies eg. countable additivity.

Note. A deeper measure theory course will approach this problem by defining the probablity function on an algebra of subsets rather than on each subset directly. Refer to tools like Caratheodory's extension theorem.