Probability as expectation

Alex Fu

2023-02-05

Probability spaces are not the only possible foundations for probability theory. For example, free probability roughly considers random variables, together with algebraic operations and the linear functional of expectation, to be the central objects of study. A probability space might be described by its "identity random variable," like in category theory. Treating simpler objects as particular cases of more "complex" constructions can often give more insight into the simpler constructions.

For instance, naïvely constructing the positive set in the Hahn–Jordan decomposition theorem using set operations fails. Instead, considering sets as indicator functions offers more freedom: after verifying Cauchyness in L^1, we find that the limiting function is none other than another indicator function, from which we convert back to a set.

The identity $\mathbb{E}(\mathbb{1}_A) = \mathbb{P}(A)$ is also useful in this manner. Expectation may seem like the more complex construction compared to probability, but it offers additional key properties like linearity. Among other things, this gives us an alternate derivation of the principle of inclusion-exclusion, now using De Morgan's laws:

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) &= \mathbb{E}(1 - (1 - \mathbb{1}_{A_1}) \cdots (1 - \mathbb{1}_{A_n})) \\ &= 1 - \left[1 - \mathbb{E}\left(\sum_{i=1}^n \mathbb{1}_{A_i} + \sum_{i<j} \mathbb{1}_{A_i}\mathbb{1}_{A_j} - \cdots + (-1)^n \prod_{i=1}^n \mathbb{1}_{A_i}\right)\right] \\ &= \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \cdots + (-1)^{n-1}\, \mathbb{P}\left(\bigcap_{i=1}^n A_i\right).\end{aligned}$$

The product $\prod_{i=1}^n (1 - \mathbb{1}_{A_i})$ expands out to the familiar 2^n terms in the principle of inclusion-exclusion. Note that $\mathbb{1}_A \cdot \mathbb{1}_B = \mathbb{1}_{A \cap B}$ is always true, regardless of independence. More importantly, this technique provides a proof for **Bonferroni's inequalities**, alternating lower and upper bounds that generalize the principle of inclusion-exclusion and Boole's inequality (subadditivity or the union bound).

Proposition 1.

Let $S_k = \sum_{i_1 < \cdots < i_k} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k})$ be the kth term in the principle of inclusion-exclusion, which states that

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

A finer statement is that the partial sums form alternating lower and upper bounds by parity:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \begin{cases} \leq \sum_{k=1}^{m}(-1)^{k-1}S_k & \text{if } m \in \{1,\ldots,n\} \text{ is odd} \\ \geq \sum_{k=1}^{m}(-1)^{k-1}S_k & \text{if } m \in \{1,\ldots,n\} \text{ is even.} \end{cases}$$

Proof. Let $T = \sum_{i=1}^{n} \mathbb{1}_{A_i}$. We observe that by the linearity of expectation,

$$S_k = \sum_{i_1<\cdots<i_k} \mathbb{E}(\mathbb{1}_{A_{i_1}} \cdots \mathbb{1}_{A_{i_k}}) = \mathbb{E}\binom{T}{k}.$$

Then, we are done by a result on the truncated sums of alternating binomial coefficients. By induction on m, we can check the following identity:

$$\sum_{k=0}^{m}(-1)^k\binom{T}{k} = (-1)^m\binom{T-1}{m} \qquad \text{for } m = 1,\ldots,T-1.$$

Note that $\binom{T}{m+1} - \binom{T-1}{m} = \binom{T-1}{m+1}$, also known as *Pascal's identity*. The number of ways to choose $m+1$ elements out of T is the sum of (the number of choices of m elements out of $T-1$) and (the number of choices of $m+1$ elements out of $T-1$), depending on a given element is included or excluded.

For $m = T$, we set $\binom{T-1}{T} = 0$ by convention. That is, $\sum_{k=0}^{T}(-1)^k\binom{T}{k} = 0$, which is apparent from the symmetry of binomial coefficients for odd T, though not immediately obvious for even T. Now, to be more explicit,

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) - \sum_{k=1}^{m}(-1)^{k-1}S_k &= \mathbb{E}\left(\mathbb{1}_A - \sum_{k=1}^{m}(-1)^{k-1}\binom{T}{k}\right) \\ &= \mathbb{E}\,\mathbb{1}_A\sum_{k=0}^{m}(-1)^k\binom{T}{k} \\ &= (-1)^m\,\mathbb{E}\,\mathbb{1}_A\binom{T-1}{m} \end{aligned}$$

is nonpositive when m is odd and nonnegative when m is even. Note that $A = \bigcup_{i=1}^{n} A_i$, and $T = 0$ when $\mathbb{1}_A = 0$.

Alternatively, we may have used the fact that S_k is the expected value of the kth elementary symmetric polynomial e_k in the indicators $\mathbb{1}_{A_1},\ldots,\mathbb{1}_{A_n}$, and $1-(1-\mathbb{1}_{A_1})\cdots(1-\mathbb{1}_{A_n})$ is the indicator of $\bigcup_{i=1}^{n} A_i$. ☐

For another application of the linearity of expectation with indicator identities, consider the following formula, which, like Proposition 1, would be difficult to derive from the first principles of probability.

Proposition 2.

The probability that exactly m events in $A_1,\ldots,A_n$ occur is

$$\sum_{k=m}^{n}(-1)^{k-m}\binom{k}{m}S_k = S_m - \binom{m+1}{m}S_{m+1} + \cdots + (-1)^{n-m}\binom{n}{m}S_n.$$

When $m = 1$, this is the usual principle of inclusion-exclusion.

Proof. Let $T = \sum_{i=1}^{n} \mathbb{1}_{A_i}$ as before, and let B be the desired event that $T = m$. Then

$$\mathbb{1}_B = \sum_{\substack{I \subseteq [n] \\ |I|=m}} \mathbb{1}\left\{ \bigcap_{i \in I} A_i \cap \bigcap_{j \notin I} A_j^c \right\} = \sum_{\substack{I \subseteq [n] \\ |I|=m}} \mathbb{1}_{\bigcap_{i \in I} A_i} \cdot \left(1 - \mathbb{1}_{\bigcup_{j \notin I} A_j}\right).$$

Now, by the principle of inclusion-exclusion, the last term is

$$\mathbb{1}_{\bigcup_{j \notin I} A_j} = \sum_{k=1}^{n-m} (-1)^{k-1} \sum_{\substack{K \subseteq [n] \setminus I \\ |K|=k}} \mathbb{1}_{\bigcap_{k \in K} A_k}.$$

Distributing the multiplication (or intersection) by $\mathbb{1}_{\bigcap_{i \in I} A_i}$, we have that

$$\mathbb{1}_B = \sum_{\substack{I \subseteq [n] \\ |I|=m}} \mathbb{1}_{\bigcap_{i \in I} A_i} + \sum_{k=1}^{n-m} (-1)^k \sum_{\substack{I \subseteq [n] \\ |I|=m}} \sum_{\substack{K \subseteq [n] \setminus I \\ |K|=k}} \mathbb{1}_{\bigcap_{i \in I \sqcup K} A_i}.$$

The first sum is over all subsets with m distinct indices. The second sum counts every subset of size $m + k$ exactly $\binom{m+k}{m}$ times, which is the number of ways to split such a subset into $I \sqcup K$ with $|I| = m$ and $|K| = k$. Finally, taking expectations on both sides,

$$\mathbb{P}(B) = \mathbb{E}(\mathbb{1}_B) = S_m + \sum_{k=1}^{n-m} (-1)^k \binom{m+k}{m} S_{m+k}.$$

□

For a final remark, if probability spaces are the spaces and random variables the morphisms in probability theory, then "probability subspaces" are given by conditional spaces $(B, \mathcal{F}|_B, \mathbb{P}(\cdot \mid B))$ for $\mathbb{P}(B) > 0$. The restricted event space $\mathcal{F}|_B$ is $\{A \cap B : A \in \mathcal{F}\}$, or equivalently $\{A \subseteq B : A \in \mathcal{F}\}$, while the conditional probability measure $\mathbb{P}(\cdot \mid B)$ is suitably normalized as $\mathbb{P}(\cdot \cap B)/\mathbb{P}(B)$. Conditioning on B is the same as restricting to outcomes where B occurs.

However, a more powerful definition is the **regular** conditional probability $\mathbb{P}(A \mid \mathcal{F}) = \mathbb{E}(\mathbb{1}_A \mid \mathcal{F})$, which contains the notions of $\mathbb{P}(A \mid X(\omega)) = \mathbb{P}(A \mid \sigma(X))(\omega)$ and $\mathbb{P}(A \mid B) = \mathbb{P}(A \mid \sigma(B)) = \mathbb{P}(A \mid \{\varnothing, B, B^c, \Omega\})$. This "probability as expectation" finds applications in Markov transition kernels, random measures, Radon spaces, etc.

■