Modern Statistics

Xiangyu Chang

March 6, 2026

Abstract

To be undated.

1 Lecture 2: Random Variables and Distributions

In this lecture, we extend the foundation of probability theory to **random variables**—the bridge between abstract sample spaces and concrete numerical quantities we can analyze. We begin with the law of total probability and Bayes' theorem, essential tools for reasoning about compound events. Then we introduce random variables, their distributions, and the cumulative distribution function (CDF), which provides a complete characterization of probabilistic behavior.

1.1 The Law of Total Probability

Definition 1.1 (Partition)**.** A collection of events $\{B_1, B_2, \ldots, B_n\}$ forms a **partition** of the sample space Ω if:

- $B_i \cap B_j = \varnothing$ for all $i \neq j$.
- $\bigcup_{i=1}^{n} B_i = \Omega$.

Theorem 1.2 (Law of Total Probability)**.** *Let $\{B_1, B_2, \ldots, B_n\}$ be a partition of Ω with $P(B_i) > 0$ for all i. Then for any event A:*

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

Proof: Using the countable additivity axiom and the definition of conditional probability, we write A as a disjoint union:

$$A = \bigcup_{i=1}^{n} (A \cap B_i).$$

Since $(A \cap B_i) \cap (A \cap B_j) = \varnothing$ for $i \neq j$, we have:

$$P(A) = P\left(\bigcup_{i=1}^{n} (A \cap B_i)\right) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A \mid B_i) P(B_i).$$

■

1.2 Bayes' Theorem

Bayes' theorem is fundamental to statistical inference. It allows us to update our beliefs about the world when we observe new evidence.

Theorem 1.3 (Bayes' Theorem)**.** *For any events $B_1, B_2, \dots$ forming a partition of Ω with $P(B_i) > 0$, and any event A with $P(A) > 0$:*

$$P(B_k \mid A) = \frac{P(A \mid B_k)P(B_k)}{\sum_{i=1}^{\infty} P(A \mid B_i)P(B_i)}. \tag{2}$$

Example 1.4 (Naive Bayes Spam Filter)**.** Suppose:

- 70% of emails are spam ($P(\text{Spam}) = 0.7$).
- 20% of emails are low priority ($P(\text{Low Priority}) = 0.2$).
- 10% of emails are high priority ($P(\text{High Priority}) = 0.1$).
- The word "essay writing" appears in:
  - 90% of spam emails ($P(\text{"essay writing"} \mid \text{Spam}) = 0.9$).
  - 1% of legitimate emails ($P(\text{"essay writing"} \mid \text{Low Priority}) = 0.01$).
  - 1% of legitimate emails ($P(\text{"essay writing"} \mid \text{High Priority}) = 0.01$).

Calculate the probability that an email containing "essay writing" is spam.
Solution:

$$\begin{aligned}
P(\text{Spam} \mid \text{"essay writing"}) &= \frac{P(\text{"essay writing"} \mid \text{Spam})P(\text{Spam})}{P(\text{"essay writing"})}, \\
P(\text{"essay writing"}) &= P(\text{e.w.} \mid \text{Spam})P(\text{Spam}) + P(\text{e.w.} \mid \text{Low Priority})P(\text{Low Priority}) \\
&\quad + P(\text{e.w.} \mid \text{High Priority})P(\text{High Priority}) \\
&= 0.9 \times 0.7 + 0.01 \times 0.2 + 0.01 \times 0.1 \\
&= 0.633, \\
P(\text{Spam} \mid \text{"essay writing"}) &= \frac{0.9 \times 0.7}{0.633} = \frac{0.63}{0.633} \approx 0.9953.
\end{aligned}$$

Bayes' theorem is fundamental to statistical inference. It allows us to update our beliefs about the world when we observe new evidence. In the example above, we start with a prior belief about how likely an email is to be spam, and then update it based on the presence of certain words—the posterior probability.

1.3 Continuity of Probability

One of the most important properties of probability measures is their continuity. This property connects the behavior of probability with limits—it tells us that probability behaves nicely as we approach limits of events.

Definition 1.5 (Monotonic Sequences of Events)**.** A sequence of events $\{A_n\}_{n=1}^{\infty}$ is:

- **Increasing** if $A_1 \subseteq A_2 \subseteq \cdots$ (denoted $A_n \uparrow A$).
- **Decreasing** if $A_1 \supseteq A_2 \supseteq \cdots$ (denoted $A_n \downarrow A$),

where $A = \lim_{n\to\infty} A_n$.

Theorem 1.6 (Continuity of Probability)**.** *For any probability measure P:*

1. *If $A_n \uparrow A$, then $\lim_{n\to\infty} P(A_n) = P(A)$.*
2. *If $A_n \downarrow A$ and $P(A_1) < \infty$, then $\lim_{n\to\infty} P(A_n) = P(A)$.*

Proof. We prove the case for increasing sequences; the decreasing case follows similarly.

Suppose $A_n \uparrow A$, i.e., $A_1 \subseteq A_2 \subseteq \cdots$ and $A = \bigcup_{n=1}^{\infty} A_n$. Define the disjoint differences $B_1 = A_1$, $B_n = A_n \setminus A_{n-1}$ for $n \geqslant 2$. Then $\{B_n\}_{n=1}^{\infty}$ are pairwise disjoint sets with $\bigcup_{k=1}^{n} B_k = A_n$ and $\bigcup_{n=1}^{\infty} B_n = A$.

By countable additivity of probability measures,

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

Similarly,

$$P(A_k) = P\left(\bigcup_{n=1}^{k} B_n\right) = \sum_{n=1}^{k} P(B_n).$$

Therefore, as $k \to \infty$, $P(A_k) \to P(A)$, that is,

$$\lim_{n\to\infty} P(A_n) = P(A).$$

For the decreasing case $A_n \downarrow A$, consider the complements $A_n^c \uparrow A^c$. By the above result for increasing sequences,

$$\lim_{n\to\infty} P(A_n^c) = P(A^c).$$

Therefore,

$$\lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} \left(P(\Omega) - P(A_n^c)\right) = P(\Omega) - P(A^c) = P(A).$$

For the decreasing case, we require $P(A_1) < \infty$ to ensure the above calculations are valid. ∎

1.4 Random Variables

So far we have studied probability at the level of events. However, in most applications we are interested in **numerical** quantities that arise from random experiments. A random variable is a function that assigns a real number to each outcome in the sample space, allowing us to work with numbers rather than abstract sets.

Definition 1.7 (Random Variable)**.** A **random variable** X is a measurable function from a probability space $(\Omega, \mathcal{F}, P)$ to the real numbers $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$:

$$X : \Omega \to \mathbb{R}.$$

For every outcome $\omega \in \Omega$, the random variable assigns a real number $X(\omega)$. The requirement of **measurability** ensures that for any Borel set $B \in \mathcal{B}(\mathbb{R})$, the preimage $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$ is an event in $\mathcal{F}$, so that we can assign probabilities to it.

Example 1.8 (Coin Toss)**.** Consider the probability space $(\Omega, \mathcal{F}, P)$ where:

- Sample space $\Omega = \{H, T\}$ represents the elementary outcomes.
- Define the **indicator random variable** $X : \Omega \to \{0, 1\}$ by:

$$X(\omega) = \begin{cases} 1, & \text{if } \omega = H \text{ (Head)}. \\ 0, & \text{if } \omega = T \text{ (Tail)}. \end{cases}$$

Example 1.9 (Stock Price of Today is Higher than Yesterday)**.** Consider a financial asset's daily closing prices over time. Define:

- **Sample space**: $\Omega = \{\omega\}_{t \in \mathbb{N}}$ where ω represents whether the stock price of today is higher than yesterday.
- **Event space**: $\mathcal{F} = \sigma(\{H, L\})$ where:
  - H, price increase.
  - L, price decrease.
- **Random variable**: $X : \Omega \to \{0, 1\}$ defined as:

$$X(\omega) = \begin{cases} 1, & \text{if } \omega \in H. \\ 0, & \text{if } \omega \in L. \end{cases}$$

Example 1.10 (Five Coin Tosses)**.** Consider the experiment of five independent coin tosses:

- **Sample space**: $\Omega = \big\{\omega = (\omega_1, \dots, \omega_5) \mid \omega_i \in \{H, T\}\big\}$ where H denotes Heads and T denotes Tails.
- **Random variable**: $X : \Omega \to \{0, 1, 2, 3, 4, 5\}$ defined as:

$$X(\omega) = \sum_{i=1}^{5} \mathbf{1}_{\{\omega_i = H\}},$$

  which counts the total number of Heads in the sequence.
- **Realization**: If $\omega = (H, H, H, H, H)$, then $X(\omega) = 5$.

1.5 From Random Variables to Probabilities

Once we have defined a random variable $X : \Omega \to \mathbb{R}$, we can ask probabilistic questions about it. For any Borel set $A \subseteq \mathbb{R}$, we write

$$P(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A\}).$$

In particular, for intervals $A = [a, b] \subseteq \mathbb{R}$, we can compute probabilities like $P(a \leqslant X \leqslant b)$.

Example 1.11 (Binomial Distribution)**.** Consider n independent coin tosses, each with probability p of landing heads. Let X be the random variable counting the number of heads observed:

- Sample space: $\Omega = \{HH \dots H, HH \dots T, \dots, TT \dots T\}$ (all 2^n possible sequences).
- If $\omega = \overbrace{HH \dots H}^{k} TT \dots T$ (with exactly k heads), then $X(\omega) = k$.

The probability of obtaining exactly k heads is given by the **binomial distribution**:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \dots, n.$$

Key Observations:

- When $p = \frac{1}{2}$, the distribution is symmetric about $n/2$.
- The probabilities sum to 1: $\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$. How to prove this?

1.6 Cumulative Distribution Function

The distribution of a random variable can be completely characterized by its Cumulative Distribution Function (CDF), which encodes all probabilistic information about the variable in a single function.

Definition 1.12 (Cumulative Distribution Function)**.** For a random variable X, its **Cumulative Distribution Function (CDF)** is the function $F_X : \mathbb{R} \to [0, 1]$ defined by:

$$F_X(x) = P(X \leqslant x) = P(\{\omega \in \Omega \mid X(\omega) \leqslant x\}).$$

Theorem 1.13 (CDF Characterization Theorem)**.** *A function $F : \mathbb{R} \to [0, 1]$ is the CDF of some random variable if and only if it satisfies the following four properties:*

1. ***Boundedness**: $0 \leqslant F(x) \leqslant 1$ for all $x \in \mathbb{R}$.*
2. ***Monotonicity**: $x_1 < x_2 \Rightarrow F(x_1) \leqslant F(x_2)$.*
3. ***Normalization**:*
$$\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to +\infty} F(x) = 1.$$
4. ***Right-continuity**: $F(x^+) = F(x)$ for all $x \in \mathbb{R}$.*

Theorem 1.14 (Properties of the CDF)**.** *For any random variable X with CDF F_X:*

1. *Point probability:* $P(X = x) = F_X(x) - F_X(x^-)$, *where* $F_X(x^-)$ *denotes the left limit.*

2. *If* F_X *is continuous at* x, *then* $P(X = x) = 0$, *and for any* $a \leqslant b$,

$$P(a \leqslant X \leqslant b) = P(a < X \leqslant b) = P(a \leqslant X < b) = P(a < X < b).$$

3. *For* $a < b$, $P(a < X \leqslant b) = F_X(b) - F_X(a)$.

4. $P(X > x) = 1 - F_X(x)$.

References