

STAT410: Intro to Probability Theory

Yuval Amit

Contents

1	Probability: The Mathematical Study of Uncertainty	2
2	Continuity of Probability Measures	3
3	Inclusion-Exclusion Principle	4
4	Intro to Random Variables	5
5	Conditional Probability	5
6	Independence	7
7	Random Variables	7
8	Expected Values	10

1. Probability: The Mathematical Study of Uncertainty

Definition

1.1

A **probability space** $(\Omega, \mathcal{F}, P)$ is composed of

- (i) The **sample space** Ω . It contains all possible outcomes of the “experiment”.
- (ii) The **event space** $\mathcal{F}$ consisting of **events**. An event is a collection of outcomes.
- (iii) The **probability measure** $P : \mathcal{F} \rightarrow \mathbb{R}$.

Example 1.2. Toss a fair coin. This has the sample space $\Omega := \{H, T\}$, the event space $\mathcal{F} = \mathbb{P}(\Omega)$, and the probability measure $P : \mathcal{F} \rightarrow \mathbb{R}$ given by

$$P(\{H\}) = 1/2, \quad P(\{T\}) = 1/2, \quad P(\emptyset) = 0, \quad P(\{H, T\}) = 1$$

Example 1.3. Roll a fair dice. This has the sample space $\Omega = \{1, \dots, 6\}$, the event space $\mathcal{F} = \mathbb{P}(\Omega)$. Note that $\mathcal{F}$ is too big to succinctly write what the probability measure P does. More on this later.

Example 1.4. Pick a number uniformly at random from $[0, 1]$. Here, $\Omega = [0, 1]$. We pretend that $\mathcal{F} = \mathbb{P}([0, 1])$, but in reality $\mathcal{F} \subsetneq \mathbb{P}(\Omega)$. Again, we can't explicitly write what P does.

Definition

1.5

We define the following axioms for the event space $\mathcal{F}$:

- 1. $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$.
- 2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
- 3. If $\{A_i\}$ is a sequence of events in $\mathcal{F}$, then $\bigcup A_i \in \mathcal{F}$.

This is the definition of a **σ -algebra**.

Remark 1.6. If $A_1, \dots, A_n \in \mathcal{F}$, then $\bigcup_{i=1}^n A_i \in \mathcal{F}$.

Proof. Let $A_i = \emptyset$ for all $i > n$. Then, we know $\bigcup_{i=1}^n A_i = \bigcup_{i=1}^\infty A_i \in \mathcal{F}$ □

Proposition

1.7

If $\{A_i\}$ is a sequence of events in $\mathcal{F}$, then $\bigcap A_i \in \mathcal{F}$.

Proof.

$$\bigcap_{i=1}^\infty A_i = \left(\bigcup_{i=1}^\infty A_i^c \right)^c \in \mathcal{F}$$

by axiom 2, 3, and De Morgan's law. □

Definition

1.8

We define the following axioms for the probability measure P :

- 1. $P(\emptyset) = 0$ and $P(\Omega) = 1$.
- 2. For any $A \in \mathcal{F}$, $P(A) \geq 0$.

3. (Countable additivity) If $\{A_i\}$ is a sequence of pairwise-disjoint events, then

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

Example 1.9. Let $\Omega = \{1, \dots, 6\}$ be the sample space for rolling a fair dice, and $\mathcal{F} = \mathbb{P}(\Omega)$. Recall that $P(\Omega) = 1$, and moreover

$$\Omega = \{1\} \sqcup \{2\} \sqcup \dots \sqcup \{6\}.$$

Thus,

$$1 = P(\Omega) = \sum_{i=1}^6 P(\{i\}).$$

By the assumption that the dice is fair, we know $P(\{i\}) = P(\{j\})$, hence $P(\{i\}) = 1/6$ for all $i \in \{1, \dots, 6\}$.

Example 1.10. Suppose $\Omega = [0, 1]$, and for any subinterval $I \subset [0, 1]$, $P(I) = \text{length of } I$. What is the probability you see a rational number? An irrational number?

Note that for any $x \in [0, 1]$, $P(\{x\}) = 0$ since $\{x\}$ can be considered a closed interval with length 0. Thus, since $\mathbb{Q}$ is countable, it follows that

$$P(\mathbb{Q} \cap [0, 1]) = 0, \quad P((\mathbb{R} \setminus \mathbb{Q}) \cap [0, 1]) = 1$$

Proposition

1.11

$$P(B^c) = 1 - P(B)$$

2. Continuity of Probability Measures

Let $B_1 \supset B_2 \supset \dots$ be a decreasing sequence of sets. Then, we write

$$\lim_{n \rightarrow \infty} B_n = B = \bigcap_{i=1}^{\infty} B_i,$$

which is the greatest lower bound, or the infimum.

Theorem

2.1

The probability measure P commutes with limits. That is, given an ascending sequence $C_1 \subset C_2 \subset \dots$,

$$P\left(\bigcup_{n=1}^{\infty} C_n\right) = P\left(\lim_{n \rightarrow \infty} C_n\right) = \lim_{n \rightarrow \infty} P(C_n)$$

Proof. Let $D_1 = C_1$, $D_2 = C_2 - C_1$, and generally $D_{i+1} := C_{i+1} - C_i$. Then, observe

$$\bigcup_{i=1}^{\infty} C_i = \bigcup_{i=1}^{\infty} D_i,$$

hence

$$P\left(\bigcup C_i\right) = P\left(\bigcup D_i\right).$$

Moreover, since the D_i 's are pairwise disjoint, one shows that

$$P(D_{i+1}) = P(C_{i+1}) - P(C_i),$$

hence

$$\sum_{i=1}^n P(D_i) = P(C_n).$$

Thus,

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} C_i\right) &= P\left(\bigcup_{i=1}^{\infty} D_i\right) = \sum_{i=1}^{\infty} P(D_i) \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n P(D_i) = \lim_{n \rightarrow \infty} P(C_n) \end{aligned}$$

as desired. $\square$

3. Inclusion-Exclusion Principle

Inclusion-Exclusion Principle

3.1

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Proof. Observe

$$P(A \cup B) = P(A \cup (B - A)) = P(A) + P(B - A).$$

Moreover,

$$P(B) = P(A \cap B) + P(B - A) \implies P(B - A) = P(B) - P(A \cap B),$$

thus

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

$\square$

Corollary

3.2

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

Proposition

3.3

$$\begin{aligned} P\left(\bigcup_{i=1}^n A_i\right) &\leq \sum_{i=1}^n P(A_i) \\ P\left(\bigcup_{i=1}^{\infty} A_i\right) &\leq \sum_{i=1}^{\infty} P(A_i) \end{aligned}$$

Proposition

3.4

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

4. Intro to Random Variables

Definition

4.1

A **random variable** is a (measurable) function $X : \Omega \rightarrow \mathbb{R}$.

Definition

4.2

The **cumulative distribution function**, abbreviated **cdf**, is

$$F_X(a) := P(X \leq a) = P(X \in [-\infty, a]).$$

Now, suppose X takes countably many values, i.e. $\{X(\omega) \mid \omega \in \Omega\}$ is countable. A **probability mass function**, abbreviated **pmt**, is a function p_X such that there's sequence of real numbers $\{x_n\}$ with $p_X(x_n) = P(X = x_n) > 0$ for all n , and $p_X(x) = P(X = x) = 0$ for all other x .

Example 4.3. Any experiment with finitely many outcomes X is called a *discrete random variable*. Then, $P_X(x)$ is the probability mass function of X . The set where $P_X(x) > 0$ is called the **support**.

For any set B , we have

$$P(x \in B) = \sum_{x \in \text{supp}(X) \cap B} p_X(x).$$

Definition

4.4

A random variable X is said to be **(absolutely) continuous** if there exists a function $f_X(x)$ such that $f_X(x) \geq 0$, and for any interval $I = [a, b]$,

$$P(x \in I) = \int_a^b f_X(x) dx$$

5. Conditional Probability

Question 5.1. Does knowing A has taken place affect our chances of seeing B ?

Definition

5.2

Let A, B be events. Then, the probability of B given A is denoted $P(B|A)$.

Theorem

5.3

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

Proof. We can give a frequentist proof with

$$P(B|A) = \lim_{n \rightarrow \infty} \frac{\frac{\#A, B \text{ both take place}}{n}}{\frac{\#A \text{ takes place}}{n}} = \frac{P(A \cap B)}{P(A)}.$$

□

Example 5.4. An experiment consists of tossing a fair coin, followed by rolling a fair D6 if you see heads, and rolling a fair D4 if you see tails. What is the probability that you see heads and two?

If $A = \text{coin is heads}$ and $B = \text{dice is two}$, then our probability is

$$P(A \cap B) = P(A) \cdot P(B|A) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}.$$

Theorem**5.5**

$P'(X) := P(X|A)$ is a probability measure. However, $P''(X) := P(B|X)$ is not a probability measure.

Law of Total Probability**5.6**

Let $A_1, \dots, A_n$ partition Ω . and $P(A_i) > 0$ for all i . Then,

$$P(B) = \sum_{i \in I} P(B|A_i)P(A_i)$$

This also holds for countable partitions.

Bayes' Theorem**5.7**

Suppose $A_1, \dots, A_n$ partition Ω , and $P(A_i) > 0$. Suppose further that B is an event with $P(B) > 0$. Then, for any $j \in \{1, \dots, n\}$,

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^n P(B|A_i)P(A_i)},$$

as desired.

Proof. By [Theorem 5.3](#),

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

and by the law of total probability

$$\frac{P(A_j \cap B)}{P(B)} = \frac{P(A_j \cap B)}{\sum_{i=1}^n P(B|A_i)P(A_i)} = \frac{P(B|A_j) \cdot P(A_j)}{\sum_{i=1}^n P(B|A_i)P(A_i)}$$

□

Example 5.8. The prevalence of a particular disease is 1%. A rapid test for this disease has a 2% false positive rate, and a 3% false negative rate. If the test comes up positive, how likely is it that you have the disease?

Solution. Let D be the event that you have the disease, and T be the event that you tested positive. Then,

$$P(D) = 0.01,$$

$$P(T|D^c) = 0.02,$$

$$P(T^c|D) = 0.03.$$

By Bayes' Theorem,

$$P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^c)P(D^c)} = \frac{(1 - 0.03) \cdot 0.01}{(1 - 0.03) \cdot 0.01 + 0.02 \cdot (1 - 0.01)} = 0.3288$$

□

6. Independence

Definition

6.1

Events A and B are said to be **independent** if $P(A \cap B) = P(A)P(B)$.

Theorem

6.2

Let A, B be events with $P(A), P(B) > 0$. If they are independent, they cannot be disjoint.

Proof. $P(A \cap B) = P(A)P(B) > 0$.

□

Theorem

6.3

For any events $A_1, \dots, A_n$ with $P(A_i) > 0$,

$$P\left(\bigcap_{i=1}^n A_i\right) = P\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right) P\left(A_{n-1} \mid \bigcap_{i=1}^{n-2} A_i\right) \cdots P(A_1).$$

Example 6.4. An experiment consists of tossing a fair coin until you see heads for the first time, at which point the experiment terminates. Let A_i be the event you see tails on the i 'th toss. Are the A_i 's independent? What is the probability we see heads for the first time on the n 'th toss?

No, the A_i 's are not independent. Next, let B_n be the event that heads is seen on the n 'th toss. Then,

$$B_n = A_1 \cap A_2 \cap \cdots \cap A_{n-1} \cap A_n^c.$$

Therefore,

$$P(B_n) = P\left(A_n^c \mid \bigcap_{i=1}^{n-1} A_i\right) \cdot \prod_{k=1}^{n-1} P\left(A_k \mid \bigcap_{i=1}^{k-1} A_i\right) = (1/2)^n.$$

Exercise 1. If A, B are independent, show that A, B^c are independent.

7. Random Variables

Definition

7.1

For a random variable X , and a (measurable) $A \subset \mathbb{R}$, we define the distribution D_X to be

$$D_X(A) = P(X \in A)$$

Now, given a random variable $X : \Omega \rightarrow \mathbb{R}$, we have

$$D_X(A) = P(X \in A) = P(\{\omega \mid X(\omega) \in A\}),$$

and

$$F_X(a) = P(X \leq a) = P(X \in (-\infty, a]).$$

Clearly, D_X determines F_X , but in fact we'll see that F_X determines D_X .

Example 7.2. Let X be the outcome of rolling a fair die. Then,

$$F_X(a) = \begin{cases} 0 & a < 1 \\ 1/6 & 1 \leq a < 2 \\ 2/6 & 2 \leq a < 3 \\ \vdots & \\ 5/6 & 5 \leq a < 6 \\ 1 & 6 \leq a. \end{cases}$$

In other words,

$$F_X(a) = \begin{cases} 0 & a < 1 \\ 1 & a \geq 6 \\ \frac{\lfloor a \rfloor}{6} & 1 \leq a < 6. \end{cases}$$

Proposition

7.3

The function $F_X : \mathbb{R} \rightarrow \mathbb{R}$ satisfies the following properties:

1. $\lim_{X \rightarrow \infty} F_X(a) = 1$ and $\lim_{X \rightarrow -\infty} F_X(a) = 0$
2. F_X is monotonically increasing
3. F_X is right continuous, i.e.

$$\lim_{x \rightarrow a^+} F_X(x) = F_X(a)$$

Definition (Discrete Random Variables)

7.4

X is called **discrete** if there is a countable $A \subset \mathbb{R}$ such that $P(X \in A) = 1$.

Then, the function $a \mapsto P(X = a)$ is called the **probability mass function** (pmf), and the probabilities of interest are computed by taking a suitable sum of the pmf.

Definition (Continuous Random Variables)

7.5

X is called **(absolutely) continuous** if there is some function $f_X(x)$ such that $f_X(x) \geq 0$, and for any interval I with endpoints $-\infty \leq a \leq b \leq \infty$,

$$P(X \in I) = \int_a^b f_X(x) dx$$

Note that we then have

$$F_X(a) = \int_{-\infty}^a f_X(x) dx$$

Definition (Identical Distributions)

7.6

Random variables X, Y are said to be **identically distributed** if $D_X = D_Y$, which holds if and only if $F_X = F_Y$.

A good way to think about identical distributions is with an example: how do you generate a random integer in $\{0, 1\}$?

One way to do it is flipping a coin, taking 0 if the coin is tails, and 1 if the coin is heads. Another way to do it is generating a random real number in $[0, 1]$, taking 0 if it's $< 1/2$, and 1 if it's $\geq 1/2$.

Let X be the coin method and Y be the real number method. Note that X and Y have completely different sample spaces:

$$\begin{aligned} X : \{H, T\} &\rightarrow \mathbb{R} & X(\{H\}) &= 1, X(\{T\}) = 0 \\ Y : [0, 1] &\rightarrow \mathbb{R} & Y(\omega) &= \begin{cases} 0 & \omega < 0.5 \\ 1 & \omega \geq 0.5. \end{cases} \end{aligned}$$

However, observe

$$F_X(a) = F_Y(a) = \begin{cases} 0 & a < 0 \\ 1/2 & 0 \leq a < 1 \\ 1 & a \geq 1. \end{cases}$$

Thus, X and Y are identically distributed, even though they're completely different (they have different sample spaces!).

Moreover, let $W = 1 - X$, i.e. $W = 1$ on tails and $W = 0$ on heads. Then, X and W are still identically distributed!

- Remark 7.7.**
1. If F_X is piecewise constant, then X is a discrete random variable. In this case, $P(X = a)$ can be computed by $F(a) - \lim_{X \rightarrow a^-} F_X(a)$
 2. If X is (absolutely) continuous, F_X is continuous
 3. If F_X is differentiable with only finitely many exceptions, then $F'_X(x) = f_X(x)$ for all points where F_X is differentiable. At other points, we may define f_X arbitrarily.

Example 7.8. Suppose X is a continuous random variable with pdf $f_X(x) = 1/2$ for $x \in [-1, 1]$.

(i) Compute $P(X > 0)$

We know

$$P(X > 0) = \int_0^{\infty} f_X(x) dx = \int_0^1 \frac{1}{2} dx = 1/2$$

(ii) Compute $P(X > 1/2 | X > 0)$.

Using conditional probability, we see that

$$P(X > 1/2 | X > 0) = \frac{P(X > 1/2 \cap X > 0)}{P(X > 0)} = \frac{1/4}{1/2} = \frac{1}{2}.$$

(iii) Compute the pdf of X^2 .

Observe that

$$P(X^2 \leq a) = P(-\sqrt{a} \leq X \leq \sqrt{a}) = \int_{-\sqrt{a}}^{\sqrt{a}} f_X(x) dx.$$

When $0 \leq a < 1$, this evaluates to $\sqrt{a}$. Thus, we see

$$F_{X^2}(a) = \begin{cases} 0 & a < 0 \\ \sqrt{a} & 0 \leq a < 1 \\ 1 & a \geq 1. \end{cases}$$

Note that this is the cdf of X^2 . Differentiating, we get

$$f_{X^2}(a) = \begin{cases} 0 & a < 0 \\ \frac{1}{2\sqrt{a}} & 0 \leq a < 1 \\ 0 & a \geq 1. \end{cases}$$

Definition

7.9

We say that a random variable X has a **uniform distribution** on $[a, b]$, denoted $X \sim u[a, b]$, if

$$f_X(x) = \frac{1}{b-a}$$

for $x \in [a, b]$.

Example 7.10. Suppose $X \sim u[0, 1]$. Find the pdf of $Y := e^X$.

Solution. If $a < 1$, then e^X doesn't attain a on $[0, 1]$, so $F_Y(a) = 0$ for $a < 1$. Similarly, when $a > e$, we know with certainty that $e^X < a$, so $F_Y(a) = 1$. Finally, if $a \in [1, e]$, then

$$F_Y(a) = P(Y \leq a) = P(e^X \leq a) = P(X \leq \log a) = P(0 \leq X \leq \log a) = \log a.$$

Thus,

$$F_Y(a) = \begin{cases} 0 & a < 1 \\ \log(a) & 1 \leq a \leq e \\ 1 & a > e, \end{cases}$$

and differentiating yields

$$f_Y(a) = \begin{cases} 0 & a \leq 1 \\ \frac{1}{a} & 1 < a < e \\ 0 & a \geq e. \end{cases}$$

□

8. Expected Values

Definition (Expected Value (Discrete))

8.1

Let X be a discrete random variable with support A . Then, the **expected value** of X , denoted $E[X]$, is

$$E[X] = \sum_{x \in A} x p_X(x) = \sum_{x \in A} x P(X = x)$$

Definition (Expected Value (Continuous))

8.2

Let X be a continuous random variable. Then, the **expected value** of X , denoted $E[X]$, is

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

Definition

8.3

We say that the expected value $E[X]$ **exists** if $E[|X|] < \infty$.

Example 8.4. Suppose $X \sim u[0, 1]$. Compute $E[X]$ and $E[e^X]$.

Solution. By definition,

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^1 x dx = \frac{1}{2}.$$

For $E[e^X]$, one method involves using the pdf we computed earlier, which gives

$$E[e^X] = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_1^e \frac{y}{y} dy = e - 1.$$

However, we may use the following general result:

$$E[g(x)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx,$$

which gives

$$E[e^X] = \int_0^1 e^x dx = e - 1.$$

□

LOTUS

8.5

Let X be a random variable with density function f_X . Then, for $g : \mathbb{R} \rightarrow \mathbb{R}$,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Linearity of Expectation

8.6

If X, Y are random variables, and a is a real number, then $E[aX + Y] = aE[X] + E[Y]$.

Markov's Inequality

8.7

Suppose $X \geq 0$ and $c \in \mathbb{R}^+$. Then

$$P(X \geq c) \leq E[X]/c,$$

Chebyshev's Inequality

8.8

Let X be a random variable with finite mean μ and finite variance σ^2 . For any $c \in \mathbb{R}^+$,

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

Proposition**8.9**

If $E[X^k]$ is finite, then $E[X^n]$ is finite for all $n \in \{1, \dots, k\}$. In particular, for Chebyshev's inequality, it's sufficient to know that σ^2 is finite.

Proof. Observe

$$E[|X^m|] = \int_{-\infty}^{\infty} |x^m| f_X(x) dx = \int_{|x^m| > 1} |x^m| f_X(x) dx + \int_{|x^m| \leq 1} |x^m| f_X(x) dx$$

by LOTUS ([Theorem 8.5](#)), for any m . Let A be the first term above when $m = k$, and B the second term when $m = k$. When $x^n > 1$, then $|x^n| f_X(x) \leq |x^k| f_X(x)$. When $|x^m| \leq 1$, then $|x^m| f_X(x) \leq f_X(x)$. Thus,

$$E[|X^n|] \leq \int_{|x^n| > 1} |x^k| f_X(x) dx + \int_{-\infty}^{\infty} f_X(x) dx \leq A + 1 < \infty,$$

so $E[X^n]$ is finite. □