AA228/CS238: Probability Review!

Houjun Liu *September 24, 2025*

1 Random Variable

random variables takes on different values with different probabilities. Each value a **random variable** take on is an **event**.

For instance, here's a random variable representing a die: X. It can takes on the following values, with the following probabilities:

$$P(X = 1) = \frac{1}{6} \tag{1}$$

$$P(X = 2) = \frac{1}{6} \tag{2}$$

$$\dots \tag{3}$$

$$P(X = 6) = \frac{1}{6} \tag{4}$$

where each assignment $X = k$ is what we refer to above as an **event**.

The set of assignments of a random variable and their associated probability is called a *distribution*: distributions "assigns probabilities to outcomes." When we say a certain random variable X is "distributed" following a distribution D, we say $X \sim D$. Semantically, we say X is a D random variable.

2 Notation Time!

$$P(X = k) \tag{5}$$

"our random variable X takes up the value k". We (including the textbook!) write it as:

$$P(x^k) \tag{6}$$

as a shorthand.

3 A Frequentist Definition

say you performed n trials, and you are wondering what the probability of a certain event $E := x^j$ is amongst those trials

$$P(x^j) = \lim_{n \to \infty} \frac{n(x^j)}{n} \tag{7}$$

"the ratio of trials that result in the event to the number of times you tried"

4 Probability Axioms

- $0 \leq P(x^j) \leq 1, \forall X, \forall j$: all probabilities are numbers between 0 and 1
- $P(x^1 \vee \ldots \vee x^n) = 1$: set of all possible outcomes must be from the sample space
- if x^a and x^b are mutually exclusive, $P(x^a) + P(x^b) = P(x^a \vee x^b)$
  - so: if $x^a \wedge x^b = F$ (there's no world in witch x^a and x^b are both true), then $P(x^a) + P(x^b) = P(x^a \vee x^b)$

5 Probabilities Correlaries

To prove the results in this part, we will use the language of set theory. However, axioms derived in this language directly translate into the logic language before and after.

5.1 Probability of Complements

Statement:

$$P(\neg x^j) = 1 - p(x^j) \tag{8}$$

Discussion: The worlds in which $\neg x^j$ is the complement of the world in which x^j is true. Let the worlds in which x^j is true be E, then, we desire:
$P(E^C) = 1 - P(E)$
Because:
We know that E and E^C are mutually exclusive, so

$$P(S) = 1 = P(E \cup E^C) \tag{9}$$
$$= P(E) + P(E^C) \tag{10}$$

So $1 - P(E) = P(E^C)$

5.2 Probability of Subsets

Statement:

$$x^a \rightarrow x^b, P(x^a) \leq P(x^b) \tag{11}$$

Discussion: The first part implies that the worlds in which x^a is true is a subset than the words in which x^b is true (because x^a implies x^b). Meaning, if E was the worlds in which x^a is true, and F is the world in which x^b is true, we desire:
if $E \subseteq F$, then $P(F) \geq P(E)$.
Recall a result from set theory: if $E \subseteq F$, $F = E \cup (E^C \cap F)$.
Then, we have:

$$P(F) = P(E \cup (E^C \cap F)) \tag{12}$$
$$= P(E) + P(E^C \cap F) \tag{13}$$

$P(E^C \cap F) \geq 0$, so:

$$P(F) \geq P(E) \tag{14}$$

5.3 Inclusion-Exclusion Principle

Statement:

$$P(x^a) + P(x^b) - P(x^a \wedge x^b) = P(x^a \vee x^b) \tag{15}$$

Discussion: Consider A be the set of worlds in which x^a is true; and B the set of worlds in which x^b is true.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{16}$$

Again, consider: $A \cup B = A \cup (A^C \cap B)$, and $B = (A \cap B) \cup (A^C \cap B)$
so:

$$\begin{aligned} P(A \cup B) &= P(A \cup (A^C \cap B)) && (17)\\ &= P(A) + P(A^C \cap B) && (18)\\ &= P(A) + P(B) - P(A \cap B) && (19) \end{aligned}$$

6 Conditional Probabilities

x: loosing contact, y: sensor failure.
"what's the probability of us loosing contact given we had a sensor failure?"

$$P(x|y) := \frac{P(x \wedge y)}{P(y)} \tag{20}$$

for simplicity we will write **AND** with a comma:

$$P(x|y) := \frac{P(x, y)}{P(y)} \tag{21}$$

multiplying:

$$P(x|y)P(y) := P(x, y) \tag{22}$$

We can keep going!

$$P(z|x, y)P(x, y) = P(z, x, y) \tag{23}$$

stick them together:

$$P(z|x, y)P(x|y)P(y) = P(z, x, y) \tag{24}$$

so, in general:

6.1 Probability Chain Rule

$$P(a^1, a^2 \ldots, a^n) = P(a^n \mid a^1, a^2 \ldots a^{n-1}) P(a^1, a^2 \ldots a^{n-1}) \tag{25}$$

6.2 Conditioning Doesn't Change Axioms

Applying a condition does not change axioms/results if its consistent
$0 \leq P(x) \leq 1$

7 Bayes Theorem

"Inference!"
it provides us a way of going from $P(x|y) \implies P(y|x)$; let y be spam, and x be emails with the word "gold" in it. It's easy to measure $P(x|y)$ (get a bunch of spam, check for the word "gold"), and by doing this we can get the more important value of $P(y|x)$ (probability of spam given "gold".)
Recall conditional probability:

$$P(x|y) := \frac{P(x,y)}{P(y)} \tag{26}$$

and the fact that $P(x, y) = P(y, x)$. so:

$$P(x|y) = \frac{P(x,y)}{P(y)} \tag{27}$$

$$= \frac{P(y,x)}{P(y)} \tag{28}$$

$$= \frac{P(y|x)P(x)}{P(y)} \tag{29}$$

8 Independence

We define $y \perp x$ if:

$$P(x|y) = P(x) \tag{30}$$

"knowing y doesn't do anything to our knowledge of x"
Now. Consider the conditional probability:

$$P(x|y)P(y) = P(x,y) \tag{31}$$

substituting our definition in:

$$P(x)P(y) = P(x,y) \tag{32}$$

if $y \perp x$.

stuff could be **conditionally** independent:

$$P(e^1, e^2|f) = P(e^1|f)P(e^2|f) \tag{33}$$

does **not** imply $e^1 \perp e^2$
conversely, $e^1 \perp e^2$ does **not** imply $e^1|f \perp e^2|f$

9 Law of Total Probability

$$P(x) = \sum_{y \in Y} P(x, y) \tag{34}$$

meaning also:

$$P(x) = \sum_{y \in Y} P(x|y)P(y) \tag{35}$$

applying this to Bayes theorem

$$P(x|y) = \frac{P(y|x)P(x)}{\sum_x P(y|x)P(x)} \tag{36}$$

10 Practice Problems

10.1 Mammogram

Conditions:

- natural occurrence of breast cancer is 8%
- mammogram results a positive in 95% in people with breast cancer
- mammogram results a positive in 7% in people without breast cancer

What's the probability that a patient has breast cancer with a positive mammogram result?
Let x be the event that the patient has breast cancer, and y is a positive mammogram result. We want $P(x|y)$.
Let's convert each of our conditions into this formalism!

- $P(x) = 0.08$
- $P(y|x) = 0.95$
- $P(y|\neg x) = 0.07$

Now, recall we want:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y|x)P(x) + P(y|\neg x)P(\neg x)} \tag{37}$$

The only thing we don't directly have $P(\neg x)$, but recall by the properties is $P(\neg x) = 1 - P(x)$. So, $P(\neg x) = 1 - 0.08 = 0.92$.
Plugging everything in:

$$P(x|y) = \frac{0.95 \cdot 0.08}{0.95 \cdot 0.08 + 0.07 \cdot 0.92} \approx 0.5413 \tag{38}$$

10.2 Monty Hall

Three doors, prize behind one, midterm behind the other two. Assume the likelihood of the prize behind each door is equivalent, and assume that the host is playing rationally.
You picked a door, and the host said another door had a midterm. Should you switch?

WLOG you picked door 1, host said door 3 had midterm.
Let's formalize this first. Let p^i be the event that i door had prize; and h^j be the event that host picks j door.
We desire to answer:

$$P(p^1|h^3) \overset{?}{\succ} P(p^2|h^3) \tag{39}$$

---

recall: $P(p^{(i)}) = \frac{1}{3}$

10.2.1 Left Case

let us consider first:

$$P(p^1|h^3) = \frac{P(p^1, h^3)}{P(h^3)} \tag{40}$$

Let us expand this out with the LoTP:

$$\frac{P(h^3|p^1)P(p^1)}{P(h^3|p^1)P(p^1) + P(h^3|p^2)P(p^2) + P(h^3|p^3)P(p^3)} \tag{41}$$

- Recall all $P(p^j) = \frac{1}{3}$
- Now, let's consider each case:
  - $P(h^3|p^1) = \frac{1}{2}$ — the host has no bias towards opening either doors 2 or 3, just not door 1
  - $P(h^3|p^2) = 1$ — the host will definitely open door 3, because they can't open your door and door 2 has the prize
  - $P(h^3|p^3) = 0$ — a rational host will not open the door that has the prize

Plugging this in:

$$P(p^1|h^3) = \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}\frac{1}{3} + 1\frac{1}{3}} = \frac{1}{3} \tag{42}$$

10.2.2 Right Case

Note that the denominator is exactly the same

$$P(p^2|h^3) = \frac{P(p^2, h^3)}{P(h^3)} \tag{43}$$

Our numerator is $P(p^2, h^3) = P(h^3|p^2)P(p^2)$. The left value is 1, and the right value is still $\frac{1}{3}$. Plugging it in:

$$P(p^2|h^3) = \frac{1\frac{1}{3}}{\frac{1}{2}\frac{1}{3} + 1\frac{1}{3}} = \frac{2}{3} \tag{44}$$

10.2.3 Conclusion

$$P(p^1|h^3) < P(p^2|h^3) \tag{45}$$

meaning

$$p^1|h^3 \prec p^2|h^3 \tag{46}$$

so we should probably switch

11 Continuous and Discrete Probabilities

11.1 Discrete Distributions

So far we have been talking about **discrete** distributions, where a random variable takes on discrete values such as dice rolls $1:6$. These distributions use a **probability mass function** (by convention uppercase P), which is written as an assignment of probabilities to values.

As a reminder:

$P(S) = 1$, where S is the sample space. Since our probability mass function specify all possible events, we should expect:

$$\sum_X P(X) = 1 \tag{47}$$

11.2 Continuous Distributions

"what's the probability of the high tomorrow at Stanford being exactly $82.9239328452°F$?"

Vanishingly unlikely. So, events in continuous distribution are formulated as an *integral* over ranges of likely outcomes. That is:

$$P(a \leq X \leq b) \tag{48}$$

if $X \sim D$ where D is continuous.

Continuous distributions are given by a **probability density function** (PDF), which defines *changes* in probabilities over a range. Integrating it results in the actual probability value. That is, for PDF $f(x)$ (by convention lowercase f), we have:

$$P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x \tag{49}$$

We often ask for events of the shape $X \leq x$ (or, the complement thereof, $X \geq x$)—"what's the chance that it will be hotter than $90°$ tomorrow? So, we often compute the **cumulative density function** (CDF) of a probability $F(x)$ (by convention uppercase f), which is defined by:

$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(z)\,\mathrm{d}z \tag{50}$$

11.3 Moments

expected value: the "mean" of the random variable:

$$E[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x \tag{51}$$

and variance:

$$Var[X] = E[X^2] - [E[X]]^2 \tag{52}$$

12 Some Useful Distributions

See slides.

12.1 Gaussian

It's the best because the **central limit theorem** exists: if you have a bunch of independent, and identical random variables, adding more of them together results in more of a Gaussian distribution. That is, for a bunch of independent X_n where all $X_j \sim X$, we have that:

$$\sum_{i=1}^{N} X_n \sim \mathcal{N}(n\mu, n\sigma^2), \text{as } n \to \infty \tag{53}$$

13 Compute a CDF!

2.1, Chapter 2 of AlgDM: Consider a continuous random variable X, which exponential distribution parameterized by λ with density $p(x|\lambda) = \lambda e^{-\lambda x}$ with nonnegative support; compute the CDF of X.
We want:

$$F(x) = \int_{-\infty}^{x} f(z)\,\mathrm{d}z \tag{54}$$

recall we are also "parameterized by λ", meaning we have some fixed, given λ. Also, we are given our $f_\lambda(x)$; recall this function has "nonnegative support", meaning that our:

$$f_\lambda(x) = 0, x < 0 \tag{55}$$

Writing it out fully, then, our PDF is:

$$f_\lambda(x) = \begin{cases} 0, x < 0 \\ p(x|\lambda) = \lambda e^{-\lambda x}, x \geq 0 \end{cases} \tag{56}$$

Plugging it in:

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f_\lambda(z)\,\mathrm{d}z && (57)\\ &= \int_{0}^{x} \lambda e^{-\lambda z}\,\mathrm{d}z && (58)\\ &= -e^{-\lambda z}|_0^x && (59)\\ &= 1 - e^{-\lambda x} && (60) \end{aligned}$$