Probability theory A1 recap

07/04/2025

For the statements below, consider $E = (\Omega, P)$ a given probability space.

Fundamentals

Baye's Theorem and LOTP

Baye's theorem states that $\forall A, B \subset \Omega$:

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

This follows directly from the **Law of Total Probability(LOTP)**:

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

Given B_i a partition of Ω.

Notice that the function $P_C : \Omega \to [0, 1]$, $P_C(A) = P(A|C)$, given $C \subset \Omega$ is also a probability in the same space E, so both Baye's theorem and LOTP assume conditional versions written in terms of P_C.

Discrete Random Variables, Indicator Random Variables

A **Discrete Random Variable** is a function $X : \Omega \to \mathbb{R}$, with $\text{Im}(X)$ a countable set. A good example is the amount of heads in 4 tosses of a fair coin:

$$X(HHHH) = 4$$
$$X(HTHH) = 3$$
$$...$$

The values X for which a random variable assumes positive values have a special name: the **support of X**

$$\text{supp}(X) := \{\varphi \in \mathbb{R} \mid P(X = \varphi) > 0\}$$

Random variables are very useful in probability, and there is a category of them so useful and ubiquituous that it has its own name:

An **Indicator Random Variable** of $A \subset \Omega$ is $I_A : \Omega \to \mathbb{R}$ defined below:

$$I_A = \begin{cases} 1 \text{ if A occurs} \\ 0 \text{ otherwise,} \end{cases}$$

This will be of particular use after we define Expected Values:

Expected Value and Variance

The expected value of a discrete random variable can be seen as a weighted sum of all possible values of X:

$$E(X) := \sum_{\varphi \in \mathbb{R}} \varphi P(X = \varphi)$$

The most useful fact about indicator random variables is that the expected value of a i.r.v is the **probability** of the event at stake:

$$E(I_A) = P(A).$$

This is **very** useful to determine hard to calculate probabilities.

Functions of real variables are useful as well and derive a famous result, known as the **Law of the Unconscious Statistician (LOTUS)**:

Let $X, Y : \Omega \to \mathbb{R}$ be random variables and $Y = f(X), f : \mathbb{R} \to \mathbb{R}$, then:

$$E(Y) = \sum_{\varphi \in \mathbb{R}} f(\varphi) P(X = \varphi)$$

The Expected Value is a linear function and has the following properties:

$$E(aX + b) = aE(X) + b$$
$$E(X + Y) = E(X) + E(Y)$$

and if X and Y are independent,

$$E(XY) = E(X)E(Y).$$

A good method to quantify the behaviour of a r.v is using its **standard deviation**:

$$\text{SD}(X) = E(|X - E(X)|)$$

and the **variance** and **mean deviation**:

$$V(X) = E([X - E(X)]^2)$$
$$\sigma(X) = \sqrt{V(X)}.$$

The variance of a r.v has some properties too:

$$V(aX + b) = a^2 V(X)$$
$$\sigma(aX + b) = |a| \, \sigma(X)$$
$$\text{SD}(aX + b) = |a| \, \text{SD}(X)$$

and if X and Y are independent:

$$V(X + Y) = V(X) + V(Y)$$

Covariance, Correlation

We will go straight to the definition:

The **Covariance** of $X, Y : \Omega \to \mathbb{R}$ is:

$$\text{Cov}(X, Y) = E([X - E(X)(Y - E(Y))])$$

This is the same as $\text{Cov}(X) = E(XY) - E(X)E(Y)$, so the idea behind this concept is to measure how both random variables change, when analyzed together. Notice that $\text{Cov}(X, Y) = 0$ if X and Y are independent, this is intuitive

Another useful concept is the **correlation coefficient:**

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)}$$

You can verify that $\rho(aX, bY) = \rho(X, Y), \forall a, b \in \mathbb{R}$, so the units used to measure X, Y are irrelevant to their correlation coefficient.

Discrete Distributions

"The key to anihilating distributions problems in exams or wherever the fuck you find them is just to remember the stories behind them, so you can morph those stories into the conditions you currently have. So take your time and get those fucking stories in your memory."

-A guy with a P.h.D that asked not to have his name written here.

Bernoulli

Story

A random variable $X : \Omega \to \mathbb{R}$ has a **Bernoulli** distribution if we can think about it as an event with two and only two possible outcomes: succes or failure. Ex: Tossing a coin, throwing a 2-sided dice (a fucking coin).

PMF, CDF, Expected Value and Variance

A random variable $X : \Omega \to \mathbb{R}$ is said to have a Bernoulli distribution if $\text{supp}(X) := \{v_1, v_2\}, v_i \in \mathbb{R}$ and X's PMF is:

$$P(X = \varphi_1) = p$$
$$P(X = \varphi_2) = 1 - p$$

We write $X \sim \text{Bern}(p)$ with parameter p and say that X describes a Bernoulli Trial, a random experiment with 2 possible outcomes: success or failure.

The Expected value and Variance of X are:

$$E(X) = \sum_{\varphi \in \mathbb{R}} \varphi P(X = \varphi) = v_1 p + v_2(1 - p)$$

$$V(X) = E(X^2) - [E(X)]^2 = \sum_{\varphi \in \mathbb{R}} \varphi^2 P(X = \varphi) - [v_1 p + v_2(1 - p)]^2$$

$$= v_1^2 p + v_2^2(1 - p) - [v_1 p + v_2(1 - p)]^2.$$

P.S: In the very common situation where $\text{supp}(X) = \{1, 0\}$, we have $E(X) = p$ and $V(X) = $ np.

Binomial

Story

A random variable $X : \Omega \to \mathbb{R}$ has a binomial distribution if we can see it as a **series** of independent Bernoulli trials of same parameter p, such as tossing multiple coins, trying a binary experiment multiple times, etc.

PMF, CDF, Expected Value and Variance

A random variable $X : \Omega \to \mathbb{R}$ is said to have the Binomial distribution if it can be decomposed as n independent and consecutive Bernouli Trials $X_i \sim \text{Bern}(p)$, das ist:

$$X = X_1 + X_2 + ... + X_n$$

So we can see that its PMF and PDF are:

$$P(X = k) = p^k(1-p)^{n-k}\binom{n}{k}$$

$$P(X \le k) = \sum_{i=1}^{k} P(X = k)$$

The Expected value and Variance of X are:

$$E(X) = \sum_{\varphi \in \mathbb{R}} \varphi P(X = \varphi) = \sum_{\varphi \in \mathbb{R}} \varphi p^{\varphi}(1-p)^{n-\varphi}\binom{n}{\varphi} = np$$

$$V(X) = E(X^2) - [E(X)]^2 = \sum_{\varphi \in \mathbb{R}} \varphi^2 P(X = \varphi) = \sum_{\varphi \in \mathbb{R}} \varphi^2 p^{\varphi}(1-p)^{n-\varphi}\binom{n}{\varphi} - (np)^2$$

$$= npq$$

Geometric

Story

Still on the Bernoulli universe, suppose we perform a bernoulli trial with parameter p (probability of sucess), and let X be the quantity of experiments performed until the first sucess (inclusive), then we say that $X : \Omega \to \mathbb{R}$ has a **geometric** distribution with parameter p.

PMF, CDF, Expected value and Variance

The PMF and CDF of $X \sim \text{Geom}(p)$ are:

$$P(X = k) = \text{Geom}(p) = p(1-p)^{k-1}$$

$$P(X \le k) = 1 - P(X > k) = 1 - (1-p)^k.$$

Its Expected Value and Variance:

$$E(X) = \sum_{\varphi \in \mathbb{R}} \varphi P(X = \varphi) = \sum_{\varphi \in \mathbb{R}} \varphi p(1-p)^{\varphi-1} = \frac{1}{p}$$

$$V(X) = E(X^2) - [E(X)]^2 = \frac{1-p}{p^2}$$

Hypergeometric

Story

If we have an urn filled with w white and b black balls, then drawing n balls out of the urn with replacement yields a $\text{Bin}\left(n, \frac{w}{w+b}\right)$ distribution for the number of white balls obtained in n trials, because the draws are independent Bernoulli trials, each with probability $\frac{w}{w+b}$ of success. If we instead sample without replacement, then the number of white balls follows a **Hypergeometric** distribution. A good example is written below:

(Communists capture-recapture). A forest has N communists. Today, m of the communists are captured, tagged, and released into the wild. At a later date, n communists are recaptured at random. Assume that the recaptured communists are equally likely to be any set of n of the communists, e.g., a communist that has been captured does not learn how to avoid being captured again (how surprising).

By the story of the Hypergeometric, the number of tagged communists in the recaptured sample has the $\text{Hypergeom}(m, N - m, n)$ distribution. The m tagged communists in this story correspond to

the white balls and the $N - m$ untagged communists correspond to the black balls. Instead of sampling n balls from the urn, we recapture n communists from the forest.

PMF, Expected value and Variance

$X : \Omega \to \mathbb{R}$, $X \sim \text{Hypergeom}(w, b, n)$ has the following:

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$$

$$E(X) = np$$

$$V(X) = npq$$

Poisson

Story

The Poisson distribution is often used in situations where we are counting the number of successes in a particular region or interval of time, and there are a large number of trials, each with a small probability of success. For example, the following random variables could follow a distribution that is approximately Poisson.

- The number of emails you receive in an hour, There are a lot of people who could potentially email you in that hour, but it is unlikely that any specifc person will actually email you in that hour.

- The number of chips in a chocolate chip cookie. Imagine subdividing the cookie into small cubes; the probability of getting a chocolate chip in a single cube is small, but the number of cubes is large.

- The number of earthquakes in a year in some region of the world. At any given time and location, the probability of an earthquake is small, but there are a large number of possible times and locations for earthquakes to occur over the course of the year.

Now we move to:

PMF, Expected Value and Variance

IF $X : \Omega \to \mathbb{R}$ is $X \sim \text{Pois}(\lambda)$ with parameter λ, then the following hold:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

$$E(X) = \sum_{\varphi \in \mathbb{R}} \varphi P(X = \varphi) = \sum_{\varphi \in \mathbb{R}} \varphi \frac{e^{-\lambda}\lambda^\varphi}{\varphi!}$$

$$= e^{-\lambda} \sum_{\varphi \in \mathbb{R}} \frac{\varphi \lambda^\varphi}{\varphi!} = \lambda e^{-\lambda} \sum_{\varphi \in \mathbb{R}} \frac{\lambda^{\varphi-1}}{(\varphi - 1)!} = \lambda e^{-\lambda} e^\lambda = \lambda.$$

$$V(X) = E(X^2) - [E(X)]^2 = \lambda(1 + \lambda) - \lambda^2 = \lambda.$$

The conclusion $E(X^2) = \lambda(1 + \lambda)$ is not trivial, but it is true.

We now proceed to continuous random variables,

Continuous Random variables

Fundamentals

A random variable $X : \Omega \to \mathbb{R}$ is said to be **continuous** if Ω is uncountable,

This is equivalent to the possible outcomes to the random experiment performed being infinite, such as choosing a **real** number in (0,1).

A continious random variable has some interesting properties, such as the PMF being constant = 0,

To see why this is true, let X be a c.r.v. We know by the naive definition of probability that $P(X = k)$ is "the occurences of k in the support of X divided by the size of the sample space". But $|\Omega| \notin \mathbb{R}!$ ($= \infty$), therefore $P(X = k) = 0, \forall k \in \mathbb{R}!$

We now proceed with new concepts and a definition:

Definition
A random variable $X : \Omega \to \mathbb{R}$ is said to be **continuous** if its CDF is differentiable.

PDF, CDF of a c.r.v
Let $X : \Omega \to \mathbb{R}$ be a c.r.v with a differentiable $F : \mathbb{R} \to [0, 1]$ CDF, analyzing $P(X = k)$ is a waste of time, instead we use the **PDF - Probability Density Function**: The density $f : \mathbb{R} \to \mathbb{R}$ in a is, for $\varepsilon \in \mathbb{R}^+$:

$$f(a) = \lim_{\varepsilon \to 0} \frac{P(a \le X \le a + \varepsilon)}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{P(X \le a + \varepsilon) - P(X \le a)}{\varepsilon} = F'(a)$$

This is rather useful because it yields an impressive result from calculus' fundamental theorem:

$$P(a \le X \le b) = F(b) - F(a) = \int_a^b f(x)dx$$

Now calculating probabilities continously has been reduced to integrating a $\mathbb{R} \to \mathbb{R}$ function, which is not so hard.

Cauchy distribution
A c.r.v $X : \Omega \to \mathbb{R}$ has the **Cauchy Distribution** if its PDF has the form:

$$f(x) = \frac{c}{1 + x^2}, c \in \mathbb{R}$$

Functions of Continuous Random variables
Given $X : \Omega \to \mathbb{R}$ a c.r.v $f(x)$ and $F(X)$ its PDF and CDF, in that order, finding the PDF and CDF of $Y : \Omega \to \mathbb{R}, Y = h(X), h : \mathbb{R} \to \mathbb{R}$ is not so hard.

See that $Y \le y \Leftrightarrow h(X) \le y$, and the right part can be solved for $X \in I$, for some interval $I \subset \mathbb{R}$, so Y's CMF is:

$$G(Y) = P(Y \le y) = P(X \in I)$$

With G in hands, it is easy to calculate $g(y) = G'(y)$.

Example:

Let $X : \Omega \to \mathbb{R}$ have a uniform distribution in $[0, 1]$ ($F(X) = x, \forall x \in [0, 1]$), calculate the PMF and CMF of $Y = \sqrt{X}$.

Solution:

We know that for $y \in (0, 1), Y \le y \Leftrightarrow \sqrt{X} \le y \Leftrightarrow X < y^2$, and since $P(X \le x) = x$;

$$G(y) = P(Y \leq y) = P(X \leq y^2) = y^2, \text{and}$$
$$g(y) = G'(y) = 2y \text{ is the PDF},$$

Now we can move to the Expected Value and Variance of a c.r.v.

Expected value and variance of a c.r.V

As in the discrete case we had $\mu = E(X) = \sum_{\varphi=-\infty}^{\infty} \varphi P(X = \varphi)$, in the continous case, the $\sum$ becomes and $\int$, and we use the fact that $P(X = k) \sim f(k)$, so:

$$\mu = E(X) = \int_{-\infty}^{\infty} \varphi f(\varphi) d\varphi.$$

Everything you know about $E(X)$ for a discrete r.v X is valid for the discrete case, just switch the $\sum$ for $\int$.

For the variance, we had $V(X) = E([X - \mu^2])$ in the discrete case, and the continuous case is:

$$V(X) = E([X - \mu]^2) = \int_{-\infty}^{\infty} (\varphi - \mu)^2 f(\varphi) d\varphi.$$