Probability I - Notes

Arkaraj Mukherjee

September 16, 2025

For a sample space Ω, event space $\mathcal{F}$ and probability P, a *random variable* is a function $X : \Omega \to \mathbb{R}$. the set of values the random variable can take, i.e. it's range is called the *support* of the RV (short form of Random Variable) often denoted by

$$\operatorname{supp} X : = \{X(\omega) | \omega \in \Omega\}$$

RVs X for which $\operatorname{supp} X$ is atmost countable i.e. countably infinite or finite are called *discrete RVs.* The probabilities of the RV taking on some values in its support are governed by the *probability mass function* (*pmf.* in short) often denoted by p, i.e.

$$\begin{aligned} p : \operatorname{supp} X &\longrightarrow [0,1] \\ x &\mapsto P(X = x) := P(X^{-1}\{x\}) \end{aligned}$$

In practice we extend this definition of p on $\operatorname{supp} X$ all the way to $\mathbb{R}$ by fixing $p(x) = 0$ for all $x \in \mathbb{R} \setminus \operatorname{supp} X$. For discrete RVs, if we ennumerate $\operatorname{supp} X = \{x_i | i = 1, 2, \ldots\}$ then from the axioms of probability we get that,

$$\sum_{i=1}^{\infty} p(x_i) = \underbrace{\sum_{i=1}^{\infty} P(X^{-1}(\{x_i\})) = P\left(\bigcup_{i=1}^{\infty} X^{-1}(\{x_i\})\right) = P(\Omega) = 1}_{\text{Why are these three equalities true?}}$$

The *cumulative distribution function* (cdf. for short) of a random variable is a function on the real line defined as,

$$\begin{aligned} F : \mathbb{R} &\to [0,1] \\ x &\mapsto P(X \leq x) := P(X^{-1}((-\infty, x])) \end{aligned}$$

This is a non-decreasing step-function with jump discontinuitites and from the definition we can see that (try proving it)

$$p(x) = F(x) - \lim_{\substack{h \to x \\ h < x}} F(h)$$

The limit on the right is sometimes known as the *left limit of F at x.* From the axioms of probability again we can see that

$$F(x) = P(X^{-1}((-\infty, x])) = P\left(\bigcup_{x_i < x} X^{-1}(\{x_i\})\right) = \sum_{x_i < x} P(X^{-1}(\{x_i\})) = \sum_{x_i < x} p(x_i)$$

You should notice that we can also deduce the equation with the left limit from this as well. The *expectation* of a RV is a weighted average of the values it takes, formally this is written as

$$\mathbb{E}[X] := \sum_{i=1}^{\infty} x_i p(x_i)$$

It may not always exist(i.e. converge). Suppose we have a RV X, then for some function f on $\operatorname{supp} X$, $Y = f(X)$ is also a RV i.e. a function $\Omega \to \mathbb{R}$ and $\operatorname{supp} Y = f(\operatorname{supp} X)$. Also,

$$\mathbb{E}\left[Y\right] = \sum_{y \in \operatorname{supp} Y} yP(f(X) = y) = \sum_{y \in \operatorname{supp} Y} \sum_{x \in f^{-1}(y)} yP(X = x)$$

$$= \sum_{y \in \operatorname{supp} Y} \sum_{x \in f^{-1}(y)} f(x)P(X = x) = \sum_{x \in \bigcup_{y \in f(\operatorname{supp} X)} f^{-1}(y)} f(x)P(X = x)$$

$$= \sum_{x \in \operatorname{supp} X} f(x)P(X = x)$$

This is a major result. The *k-th moment* of a RV is defined as $\mathbb{E}\left[X^k\right] := \mathbb{E}\left[g(X)\right]$ where $g : x \mapsto x^k$. Let X, Y be RVs on the same sample space, then we see that when we define $(X + Y)(\omega) := X(\omega) + Y(\omega)$ then,

$$\mathbb{E}\left[X + Y\right] = \sum_{x \in \operatorname{supp} X} \sum_{y \in \operatorname{supp} Y} (x + y)P(X = x \wedge Y = y)$$

$$= \sum_{x \in \operatorname{supp} X} x \sum_{y \in \operatorname{supp} Y} P(X = x \wedge Y = y) + \sum_{y \in \operatorname{supp} Y} y \sum_{x \in \operatorname{supp} X} P(X = x \wedge Y = y)$$

$$= \sum_{x \in \operatorname{supp} X} xP(X = x) + \sum_{y \in \operatorname{supp} Y} yP(Y = y) = \mathbb{E}\left[X\right] + \mathbb{E}\left[Y\right]$$

and clearly for constant c we have $\mathbb{E}\left[cX\right] = c\mathbb{E}\left[X\right]$ thus expectation is linear. For subsets $A \subseteq \Omega$ we define an *indicator RV* I_A as the RV that is one on A and 0 everywhere else, this clearly implies that $\mathbb{E}\left[I_A\right] = P(A)$. The *mean absolute deivation (MAD)* is defined as $\mathbb{E}\left[|X - \mathbb{E}\left[X\right]|\right]$, the *variance* is defined as $\operatorname{Var}(X) := \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right]$ and the *standard deviation* is $\sqrt{\operatorname{Var}(X)}$. Clearly $\operatorname{Var}(X) \geq 0$ and $\operatorname{Var}(X) = \mathbb{E}\left[X\right]^2 - \mathbb{E}\left[X^2\right]$ thus, $\mathbb{E}\left[X\right]^2 \geq \mathbb{E}\left[X^2\right]$. Also $\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X)$. The *degenerate probability distribution* is where for some $x_o \in \mathbb{R}$ we have $\operatorname{supp} X = \{x_o\}$, for this $p(x) = \delta_{x,x_o}$ and $F(x) = 0$ if $x < x_o$ and 1 otherwise, $\mathbb{E}\left[X\right] = x_o$ and $\operatorname{Var}(X) = 0$. The *discrete uniform* distribution is where $\operatorname{supp} X = \{a, a+1, \ldots, b\}$ for some $a \leq b$ where we have $p(x) = 1/|\operatorname{supp} X| = 1/(b-a+1)$ on its support and zero elsewhere. $F(x)$ is zero for $x < a$, $|(-\infty, x] \cap \operatorname{supp} X|/|\operatorname{supp} X| = (\lfloor x \rfloor - a + 1)/(b - a + 1)$ on $[a, b]$ and 1 elsewhere. Here, $\mathbb{E}\left[X\right] = (a+b)/2$ and $\operatorname{Var}(X) = \operatorname{Var}(x - (a-1)) = ((b-a+1)^2 - 1)/12 \asymp |b-a|^2$. For some $p \in [0, 1]$ the *bernoulli distribution*, $X \sim \operatorname{Ber}(p)$ is where $\operatorname{supp} X = \{0, 1\}$ with $p(1) = p$ and $p(0) = 1 - p =$ (usually written as q), $F(x) = 0$ if $x < 0$, q on $[0, 1)$ and 1 elsewhere. The *binomial distribution*, $X \sim \operatorname{Bin}(n, p)$ for $n \in \mathbb{N}$ and $p \in [0, 1]$ is where $\operatorname{supp} X = \{0, 1, \ldots, n\}$ and $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ on its support and zero elsewhere but here $F(x)$ doesn't have a closed form. We say that two RVs X, Y are independent and write $X \amalg Y$ iff for all x, y $P(X = x \wedge y = y) = P(X = x)P(Y = y)$. Now for discrete RVs X, Y such that $X \amalg Y$ we can see that for any $A \subseteq \operatorname{supp} X$ and $B \subseteq \operatorname{supp} Y$ we have that,

$$P(X \in A \wedge Y \in B) = P\left(\bigcup_{x \in A, y \in B} X^{-1}(\{x\}) \cap Y^{-1}(\{y\})\right) = \sum_{x \in A} \sum_{y \in B} P(X = x \wedge Y = y)$$

$$= \sum_{x\in A}\sum_{y\in B} p_X(x)p_Y(y) = \left(\sum_{x\in A} p_X(x)\right)\cdot\left(\sum_{y\in B} p_Y(y)\right) = P(X\in A)\cdot P(Y\in B)$$

using usual theorems from absolute convergence of $\Sigma_{x\in A}p_X(x)$ etc. This tells us that if $X \amalg Y$ then any events defined with X, Y are also independent. Now we see that if (ignore the fact that the sum of random variables isn't defined for this to work yet, this course sucks) $X_1, \ldots, X_n \sim \mathrm{Ber}(p)$ are independent then $Y = \Sigma X_i \sim \mathrm{Bin}(n,p)$. For a proof,we can see that Y has support $\{0,\ldots,n\}$ with $p_Y(k) = P(\text{k of the } X_i\text{'s are 1 and the rest are zero}) = \binom{n}{k}p^k(1-p)^{n-k}$. Now if $X \sim \mathrm{Bin}(n,p)$ and $Y \sim \mathrm{Bin}(m,p)$ are independent then we have that, $X+Y \sim \mathrm{Bin}(m+n,p)$. This can be done in two ways, take independent bernoulli RVs $Z_1,\ldots,Z_{m+n} \sim \mathrm{Ber}(p)$ and then, $X = \sum_{i=1}^n Z_i$ and $Y = \sum_{i=1+1}^m Z_i$ giving us that $X+Y = \sum_{i=1}^{n+m} Z_i \sim \mathrm{Bin(n, p)}$. Apart from this another way is to see that,

$$p_{X+Y}(k) = \sum_{j\geq 0} P(X=j \wedge Y=k-j) = \sum_{j\geq 0} p_X(j)p_Y(k-j)$$

$$= \sum_{j\geq 0}\binom{n}{j}\binom{m}{k-j}p^{j+(k-j)}(1-p)^{n-j+(m-(k-j))}$$

$$= \sum_{j\geq 0}\binom{n}{j}\binom{m}{k-j}p^k(1-p)^{n+m-k} = [z^k]\left((1+z)^m(1+z)^n\right)p^k(1-p)^{n+m-k}$$

$$= \binom{n+m}{k}p^k(1-p)^{n+m-k}$$

where we have used the convention that $\binom{a}{b} = 0$ for $a < b$ and arrive at the same conclusion. The *poisson* distribution, $X \sim \mathrm{Pois}(\lambda)$ is when $\operatorname{supp} X = \{0,1,\ldots\}$ and $p_X(k) = e^{-\lambda}\lambda^k/k!$ for k in the support and zero elsewhere for some fixed positive number λ. This has expected value λ, variance also λ, the cdf doesn't have a closed form being, $F_X(x) = \sum_{0\leq k\leq\lfloor x\rfloor} e^{-\lambda}\lambda^k/k!$. It can be shown that the pmf increases upto $\lceil\lambda\rceil$ and then decreases by considering the ratio $p_X(m)/p_X(m-1) = \lambda/m$. The poisson distribution is *right skewed.* We also see that, if $X_n \sim \mathrm{Bin}(n,\lambda/n)$ i.e. $\mathbb{E}[X_n] = \lambda$ then $X_n \overset{\text{w.r.t. pmf}}{\approx} X \sim \mathrm{Pois}(\lambda)$. Now we show that for all $k \ll n$ we have that

$$\binom{n}{k}(\lambda^k/n^k)(1-\lambda/n)^{n-k} = p_{X_n}(k) \longrightarrow p_X(k) = e^{-\lambda}\lambda^k/k!$$

which is true as for $n \gg k$ we see that $\binom{n}{k} \sim n^k/k!$ and also the fact that $(1-\lambda/n)^{n-k} \approx e^{-\lambda}$. We see that if $X \sim \mathrm{Pois}(\lambda)$ and $Y \sim \mathrm{Pois}(\mu)$ then, $X+Y \sim \mathrm{Pois}(\lambda+\mu)$. This stems from the following algebraic manipulation:

$$p_{X+Y}(k) = \sum_{m+n=k} p_X(m)p_Y(n) = \sum_{m=0}^{k}\frac{e^{-\lambda-\mu}\lambda^m\mu^{k-m}}{m!(k-m)!}$$

$$= \frac{e^{-\lambda-\mu}}{k!}\sum_{m=0}^{k}\binom{k}{m}\lambda^m\mu^{k-m} = \frac{e^{-(\lambda+\mu)}(\lambda+\mu)^k}{k!}$$

The geometric distribution, $X \sim \mathrm{Geom}(p)$ is where $\operatorname{supp} X = \{1,2,\ldots\}$ and $p\in[0,1)$ with $p_X(k) = p(1-p)^{k-1}$, The RV X is the number of trials needed for the first success in an experiment where the probability of success is p. We easily see that $F_X(x) = 1-(1-p)^{\lfloor x \rfloor}$, $\mathbb{E}[X] = 1/p$ and $\mathrm{Var}(X) = (1-p)/p^2$. The negative binomial distribution, $X \sim NB(r,p)$ having the same support and setup as the geometric distribution (which is a special case of this) but the RV here outputs the number of trials needed for the r-th success. We clearly see that $p_X(k) = \binom{k-1}{r-1}p^r(1-p)^{k-r}$. The moments are given by,

$$\mathbb{E}[X^m] = r \cdot \left(\frac{p}{1-p}\right)^r \left((xD)^{(m-1)}\left(\frac{x^r}{(1-x)^{r+1}}\right)\right)\Bigg|_{x=1-p}$$

which we can easily get by considering the function $\mathfrak{g}(x) := \sum_{k\geq 0}\binom{k}{r}x^k = x^r/(1-x)^{r+1}$ which we can get by the linearity of the derivative and uniform convergence results. From all of this we also get that $\mathrm{Var}(X) = r(1-p)/p^2$. Now if $X_1, \ldots, X_r \overset{\text{i.i.d.}}{\sim} \text{Geom}(p)$ then, $Y = \Sigma_i X_i \sim \text{NB}(r,p)$. For thr proof,

$$p_Y(k) = \sum_{\substack{x_1+\ldots+x_r=k \\ \forall i, x_i\in\mathbb{N}}} P\left(X_1 = x_1 \text{ and} \ldots \text{and } X_r = x_r\right)$$

$$\sum_{\substack{x_1+\ldots+x_r=k \\ \forall i, x_i\in\mathbb{N}}} \prod_{i=1}^{r} P(X_i = x_i) = \sum_{x_1+\ldots+x_r=k \forall i, x_i\in\mathbb{N}} \prod_{i=1}^{r} p(1-p)^{x_i-1}$$

$$= \sum_{x_1+\ldots+x_r=k \forall i, x_i\in\mathbb{N}} p^r(1-p)^{k-r} = \binom{k-1}{r-1}p^r(1-p)^{k-r}$$

and we are done. The *hypergeometric* distribution, $X \sim \text{Hypergeom}(N,m,n)$ with $N \geq m, n$ is where the random variable outputs the number of white balls in a sample when we draw n balls without replacement from an urn containing m white balls and $N-m$ black balls, i.e. N total balls. Here we must have that $\operatorname{supp} X = \{\max\{n-(N-m), 0\}, \ldots, \min\{n,m\}\}$ and for $k \in \operatorname{supp} X$, $p_X(k) = \binom{m}{k}\binom{N-m}{n-k}/\binom{N}{n}$. This is different from the binomial distribution in the sense that here the 'p' changes whereas in the binomial distribution it was fixed and infact if we did replace things here then we would have $X \sim \text{Bin}(N, m/N)$. But if $N \gg n$ we see that not replacing things affects 'p' very slightly and we can neglect the change to write, $\text{Hypergeom}(N,m,n) \overset{\text{w.r.t. pmf}}{\approx} \text{Bin}(n, m/N)$. We can also prove that, if $X \sim \text{Hypergeom}(N,m,n)$ and $Y \sim \text{Hypergeom}(m-1, n-1, N-1)$ then,

$$\mathbb{E}\left[X^k\right] = \frac{mn}{N} \cdot \mathbb{E}\left[(Y+1)^k\right]$$

by writing $i\binom{m}{i}$ as $m\binom{m-1}{i-1}$. The probability is continuous from above meaning that for any increasing sequence of events $\{E_n\}$ i.e. $E_i \subseteq E_{i+1}$ for all $i \in \mathbb{N}$ we have,

$$\lim_{n\to\infty} P(E_n) = P\left(\lim_{n\to\infty} E_n\right)$$

where in this case $\lim_{n\to\infty} E_n := \cup_{n=1}^{\infty} E_n$ and for decreasing sequences $E_i \supseteq E_{i+1}$ the union becomes an intersection and it is identical, for a proof we consider the events $F_{i+1} = E_{i+1} \backslash E_i$ for $i \geq 1$ and $F_1 = E_1$. for the increasing case and take the complements of E_i for the decreasing case and use the result from the increasing case. We now prove some properties for $F_X(x)$,the c.d.f. Firstly it is clearly non decreasing as $x > y \implies \{\omega \in \Omega : X(\omega) \leq y\} \subseteq \{\omega \in \Omega : X(\omega) \leq x\}$. Now we claim that $\lim_{x\to\infty} F_X(x) = 1$ To prove this we pass to the sequential criterion and consider sequences all sequences $a_n \longrightarrow +\infty$. Firstly we clearly see that $\lim_{n\to\infty} F_X(n) = 1$ using the events $\{X \leq n\}$. Now by the definition of the limits we see that for all $\varepsilon > 0$ there exists nteger K such that for all $n > K, |F_X(n) - 1| < \varepsilon \implies 1 - \varepsilon < F_X(K+1) \leq 1$ and also as $a_n \to \infty \implies$ there exists some integer M such that for all $m > M, a_m > K + 1 \implies 1 - \varepsilon \leq F_X(K+1) \leq F_X(a_m) \leq 1 \implies |1 - F_X(a_m)| \leq \varepsilon$ we can conclude using the definition of a limit that $\lim_{n\to\infty} F(a_n) = 1$ and as a_n was arbitrary, by the sequential criterion,

$$\lim_{x\to\infty} F_X(x) = 1$$

And by similar arguements we can show that, $F_X(x+) = F_X(x)$ i.e. it is right continous.

continous random variables X are those RVs such that there exists a function f such that,

$$P(X^{-1}(S)) = \int_S f(x)dx$$

where S belongs to the sigma algebra generated by the intervals on $\mathbb{R}$ and f is riemann integrable on them accordingly. In this course we will only ever work with S being an interval itself and seldom call to things like

$$\{x\} = \bigcup_{n=1}^{\infty} (x - 1/n, x + 1/n)$$

to get things of form $P(X = a) = 0$ and also, $P(X \in A) = 0$ for all countable A,etc.