EECS126 Course Notes

Anmol Parande

Spring 2021 - Professor Thomas Courtade

Disclaimer: These notes reflect 126 when I took the course (Spring 2021). They may not accurately reflect current course content, so use at your own risk. If you find any typos, errors, etc, please raise an issue on the GitHub repository.

Contents

1 Introduction to Probability **2**
1.1 Conditional Probability . . . 2
1.2 Independence . . . 3

2 Random Variables and their Distributions **3**
2.1 Distributions . . . 4
2.2 Properties of Distributions . . . 6
2.2.1 Expectation . . . 6
2.2.2 Variance . . . 7
2.3 Common Distributions . . . 8
2.3.1 Discrete Distributions . . . 8
2.3.2 Continuous Distributions . . . 9

1 Introduction to Probability

Definition 1 *A probability space is a triple $(\Omega, \mathcal{F}, P)$ where Ω is a set of objects called the sample space, $\mathcal{F}$ is a family of subsets of Ω called events, and the probability measure $P : \mathcal{F} \to [0, 1]$.*

One key assumption we make is that $\mathcal{F}$ is a σ-algebra containing Ω, meaning that countably many complements, unions, and intersections of events in $\mathcal{F}$ are also events in $\mathcal{F}$. The probability measure P must obey **Kolmogorov's Axioms**.

1. $\forall A \in \mathcal{F},\ P(A) \geq 0$
2. $P(\Omega) = 1$
3. If $A_1, A_2, \dots \in \mathcal{F}$ and $\forall i \neq j,\ A_i \bigcap A_j = \emptyset$, then $P\left(\bigcup_{i\geq 1} A_i\right) = \sum_{i\geq 1} P(A_i)$

We choose Ω and $\mathcal{F}$ to model problems in a way that makes our calculations easy.

Theorem 1

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

Theorem 2 (Inclusion-Exclusion Principle)

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \left(\sum_{1\leq i_1 < \cdots < i_k \leq n} P(A_{i_1} \cap \cdots \cap A_{i_k}) \right)$$

Theorem 3 (Law of Total Probability) *If $A_1, A_2, \cdots$ partition Ω (i.e A_i are disjoint and $\cup A_i = \Omega$), then for event B,*

$$P(B) = \sum_i P(B \cap A_i)$$

1.1 Conditional Probability

Definition 2 *If B is an event with $P(B) > 0$, then the conditional probability of A given B is*

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

Intuitively, conditional probabilty is the probability of event A given that event B has occurred. In terms of probability spaces, it is as if we have taken $(\Omega, \mathcal{F}, P)$ and now have a probabilty measure $P(\cdot|C)$ belonging to the space $(\Omega, \mathcal{F}, P(\cdot|C))$.

Theorem 4 (Bayes Theorem)

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

1.2 Independence

Definition 3 *Events A and B are independent if $P(A \cap B) = P(A)P(B)$*

If $P(B) > 0$, then A, B are independent if and only if $P(A|B) = P(A)$. In other words, knowing B occurred gave no extra information about A.

Definition 4 *If A, B, C with $P(C) > 0$ satisfy $P(A \cap B|C) = P(A|C)P(B|C)$, then A and B are conditionally independent given C.*

Conditional independence is a special case of independence where A and B are not necessarily independent in the original probability space which has the measure P, but are independent in the new probability space conditioned on C with the measure $P(\cdot|C)$.

2 Random Variables and their Distributions

Definition 5 *A random variable is a function $X : \Omega \to \mathbb{R}$ with the property $\forall \alpha \in \mathbb{R}$, $\{\omega \in \Omega : \ X(\omega) \leq \alpha\} \in \mathcal{F}$.*

The condition in definition 5 is necessary to compute $P(X \leq \alpha)$, $\forall \alpha \in \mathbb{R}$. This requirement also let us compute $P(X \in B)$ for most sets by leveraging the fact that $\mathcal{F}$ is closed under complements, unions, and intersections. For example, we can also compute $P(X > \alpha)$ and $P(\alpha < X \leq \beta)$. In this sense, the property binds the probability space to the random variable.

definition 5 also implies that random variables satisfy particular algebraic properties. For example, if X, Y are random variables, then so are $X+Y, XY, X^p, \lim_{n\to\infty} X_n$, etc.

Definition 6 *A discrete random variable is a random variable whose codomain is countable.*

Definition 7 *A continuous random variable is a random variable whose codomain is the real numbers.*

Although random variables are defined based on a probability space, it is often most natural to model problems without explicitly specifying the probability space. This works so long as we specify the random variables and their distribution in a "consistent" way. This is formalized by the so-called Kolmogorov Extension Theorem but can largely be ignored.

2.1 Distributions

Roughly speaking, the distribution of a random variable gives an idea of the likelihood that a random variable takes a particular value or set of values.

Definition 8 *The probability mass function (or distribution) of a random variable X is the frequency with which X takes on different values.*

$$p_X : \mathcal{X} \to [0,1] \text{ where } \mathcal{X} = range(X), \qquad p_X(x) = Pr\left\{X = x\right\}.$$

Note that $\sum_{x\in\mathcal{X}} p_X(x) = 1$ since $\bigcap_{x\in\mathcal{X}}\{w : X(w) = x\} = \Omega$.

Continuous random variables are largely similar to discrete random variables. One key difference is that instead of being described by a probability "mass", they are instead described by a probability "density".

Definition 9 *The probability density function (distribution) of a continuous random variable describes the density by which a random variable takes a particular value.*

$$f_X : \mathbb{R} \to [0,\infty) \text{ where } \int_{-\infty}^{\infty} f_X(x)dx = 1 \text{ and } Pr\left\{X \in B\right\} = \int_B f_X(x)dx$$

Observe that if a random variable X is continuous, then the probability that it takes on a particular value is zero.

$$\Pr\left\{X = x\right\} = \lim_{\delta\to 0} \Pr\left\{x \le X \le x + \delta\right\} = \lim_{\delta\to 0} \int_x^{x+\delta} f_X(u)du = \int_x^x f_X(u)du = 0$$

Definition 10 *The cumulative distribution function (CDF) gives us the probability of a random variable X being less than or equal to a particular value.*

$$F_X : \mathbb{R} \to [0,1], \quad F_X(x) = Pr\left\{X \le x\right\}$$

Note that by the Kolomogorov axioms, F_X must satisfy three properties:

1. F_X is non-decreasing.
2. $\lim_{x \to 0} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.
3. F_X is right continuous.

It turns out that if we have any function F_X that satisfies these three properties, then it is the CDF of some random variable on some probability space. Note that $F_X(x)$ gives us an alternative way to define continuous random variables. If $F_X(x)$ is absolutely continuous, then it can be expressed as

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

for some non-negative function $f_X(x)$, and this is the PDF of a continuous random variable.

Often, when modeling problems, there are multiple random variables that we want to keep track of.

Definition 11 *If X and Y are random variables on a common probability space $(\Omega, \mathcal{F}, P)$, then the joint distribution (denoted $p_X Y(x, y)$ or $f_{XY}(x, y)$ describes the frequencies of joint outcomes.*

Note that it is possible for X to be continuous and Y to be discrete (or vice versa).

Definition 12 *The marginal distribution of a joint distribution is the distribution of a single random variable.*

$$p_X(x) = \sum_y p_{XY}(x, Y = y), \qquad f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)dy$$

Definition 13 *Two random variables X and Y are independent if their joint distribution is the product of the marginal distributions.*

Just like independence, we can extend the notion of conditional probability to random variables.

Definition 14 *The conditional distribution of X given Y captures the frequencies of X given we know the value of Y.*

$$p_{X|Y}(x|y) = \frac{P_{XY}(x, y)}{p_Y(y)}, \qquad f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

2.2 Properties of Distributions

2.2.1 Expectation

Definition 15 *The expectation of a random variable describes the center of a distribution,*

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x p_X(x), \quad \mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

provided the sum or integral converges.

Expectation has several useful properties. If we want to compute the expectation of a function of a random variable, then we can use the law of the unconscious statisitician.

Theorem 5 (Law of the Unconscious Statistician)

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) p_X(x), \quad \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Another useful property is its linearity.

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y], \ \forall a, b \in \mathbb{R}.$$

Sometimes it can be difficult to compute expectations directly. For disrete distributions, we can use the tail-sum formula.

Theorem 6 (Tail Sum) *For a non-negative integer random variable,*

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} Pr\{X \geq k\}.$$

When two random variables are independent, expectation has some additional properties.

Theorem 7 *If X and Y are independent, then*

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

We can apply expectations to conditional distributions as well.

Definition 16 *The conditional expectation of a conditional distribution is given by*

$$\mathbb{E}\left[X|Y=y\right]=\sum_{x\in\mathcal{X}}xp_{X|Y}(x|y),\quad \mathbb{E}\left[X|Y=y\right]=\int_{-\infty}^{\infty}xf_{X|Y}(x,y)dx$$

Notice that $\mathbb{E}\left[X|Y\right]$ is a function of the random variable Y, meaning we can apply theorem 5.

Theorem 8 (Tower Property) *For all functions f,*

$$\mathbb{E}\left[f(Y)X\right]=\mathbb{E}\left[f(Y)\mathbb{E}\left[X|Y\right]\right]$$

If we apply theorem 8 to the function $f(Y)=1$, then we can see that $\mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right]=\mathbb{E}\left[X\right]$.

2.2.2 Variance

Definition 17 *The variance of a discrete random variable X describes its spread around the expectation and is given by*

$$Var\left(X\right)=\mathbb{E}\left[\left(X-\mathbb{E}\left[X\right]\right)^2\right]=\mathbb{E}\left[X^2\right]-\mathbb{E}\left[X\right]^2.$$

Theorem 9 *When two random variables X and Y are independent, then*

$$Var\left(X+Y\right)=Var\left(X\right)+Var\left(Y\right).$$

Definition 18 *The covariance of two random variables describes how much they depend on each other and is given by*

$$Cov\left(X,Y\right)=\mathbb{E}\left[\left(X-\mathbb{E}\left[X\right]\right)\left(Y-\mathbb{E}\left[Y\right]\right)\right].$$

If $Cov\left(X,Y\right)=0$ then X and Y are uncorrelated.

Definition 19 *The correlation coefficient gives a single number which describes how random variables are correlated.*

$$\rho(X,Y)=\frac{Cov\left(X,Y\right)}{\sqrt{Var\left(X\right)}\sqrt{Var\left(Y\right)}}.$$

Note that $-1\leq\rho\leq 1$.

2.3 Common Distributions

2.3.1 Discrete Distributions

Definition 20 *X is uniformly distributed when each value of X has equal probability.*

$$X \sim \textit{Uniform}(\{1, 2, \cdots, n\}) \implies p_X(x) = \begin{cases} \frac{1}{n} & x = 1, 2, \cdots, n, \\ 0 & \textit{else.} \end{cases}$$

Definition 21 *X is a Bernoulli random variable if it is either 0 or 1 with $p_X(1) = p$.*

$$X \sim \textit{Bernoulli}(p) \implies p_X(x) = \begin{cases} 1-p & x = 0, \\ p & x = 1, \\ 0 & \textit{else.} \end{cases}$$

$$\mathbb{E}[X] = p \qquad Var(X) = (1-p)p$$

Bernoulli random variables are good for modeling things like a coin flip where there is a probability of success. Bernoulli random variables are frequently used as indicator random variables $\mathbb{1}_A$ where

$$\mathbb{1}_A = \begin{cases} 1 & \text{if A occurs,} \\ 0 & \text{else.} \end{cases}$$

When paired with the linearity of expectation, this can be a powerful method of computing the expectation of something.

Definition 22 *X is a Binomial random variable when*

$$X \sim \textit{Binomial}(n, p) \implies p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, \cdots, n \\ 0 & \textit{else.} \end{cases}$$

$$\mathbb{E}[X] = np \qquad Var(X) = np(1-p)$$

A binomial random variable can be thought of as the number of successes in n trials. In other words,

$$X \sim \text{Binomial}(n, p) \implies X = \sum_{i=1}^{n} X_i, \quad X_i \sim \text{Bernoulli}(p).$$

By construction, if $X \sim \text{Binomial}(n, p)$ and $Y \sim \text{Binomial}(m, p)$ are independent, then $X + Y \sim \text{Binomial}(m + n, p)$.

Definition 23 *A Geometric random variable is distributed as*

$$X \sim \textit{Geom}(p) \implies p_X(x) = \begin{cases} p(1-p)^{x-1} & x = 1, 2, \cdots \\ 0 & \textit{else}. \end{cases}$$

$$\mathbb{E}\left[X\right] = \frac{1}{p} \qquad Var\left(X\right) = \frac{1-p}{p^2}$$

Geometric random variables are useful for modeling the number of trials required before the first success. In other words,

$$X \sim \text{Geom}(p) \implies X = \min\{k \geq 1 : X_k = 1\} \text{ where } X_i \sim \text{Bernoulli}(p).$$

A useful property of geometric random variables is that they are memoryless:

$$\Pr\left\{X = K + M | X > k\right\} = \Pr\left\{X = M\right\}.$$

Definition 24 *A Poisson random variable is distributed as*

$$X \sim Poisson(\lambda) \implies p_X(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & x = 0, 1, \cdots \\ 0 & \textit{else.} \end{cases}$$

$$\mathbb{E}\left[X\right] = \lambda$$

Poisson random variables are good for modeling the number of arrivals in a given interval. Suppose you take a given time interval and divide it into n chunks where the probability of arrival in chunk i is $X_i \sim \text{Bernoulli}(p_n)$. Then the total number of arrivals $X_n = \sum_{i=1}^{n} X_i$ is distributed as a Binomial random variable with expectation $np_n = \lambda$. As we increase n to infinity but keep λ fixed, we arrive at the poisson distribution.

A useful fact about Poisson random variables is that if $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ are independent, then $X + Y \sim \text{Poisson}(\lambda + \mu)$.

2.3.2 Continuous Distributions

Definition 25 *A continuous random variable is uniformly distributed when the pdf of X is constant over a range.*

$$X \sim \textit{Uniform}(a, b) \implies f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \textit{else.} \end{cases}$$

The CDF of a uniform distribution is given by

$$F_X(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & x \in [a, b) \\ 1, & x \geq b. \end{cases}$$

Definition 26 *A continuous random variable is exponentially distributed when its pdf is given by*

$$X \sim \mathit{Exp}(\lambda) \implies f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & \mathit{else.} \end{cases}$$

Exponential random variables are the only continuous random variable to have the memoryless property:

$$\Pr\{X > t + s | X > s\} = \Pr\{X > t\}, \quad t \geq 0.$$

The CDF of the exponential distribution is given by

$$F_X(x) = \lambda \int_0^x e^{-\lambda u} du = 1 - e^{-\lambda x}$$

Definition 27 *X is a Gaussian Random Variable with mean μ and variance σ^2 (denoted $X \sim \mathcal{N}(\mu, \sigma^2)$) if it has the PDF*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

The standard normal is $X \sim \mathcal{N}(0, 1)$, and it has the CDF

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{\frac{-u^2}{2}} du$$

There is no closed from for $\Phi(x)$. It turns out that every normal random variable can be transformed into the standard normal (i.e $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$). Some facts about Gaussian random variables are

1. If $X \sim \mathcal{N}(\mu_x, \sigma_x^2),\ Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ are independent, then $X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

2. If X, Y are independent and $(X + Y), (X - Y)$ are independent, then both X and Y are Gaussian with the same variance.