ChatGPT

# Probability Space

A **probability space** formalizes the setup of a random experiment. It consists of a *sample space* $\Omega$ (the set of all possible outcomes), a collection of events (subsets of $\Omega$) forming a $\sigma$-algebra $\sigma(\Omega)$, and a probability measure $P$. The sample space and event definitions allow us to talk about occurrences of outcomes (e.g. $\Omega$ could be all outcomes of rolling two dice [1]). The probability measure $P$ assigns a number in $[0,1]$ to each event with the axioms: $P(\Omega)=1$, $P(A)\ge0$ for any event $A$, and countable additivity on disjoint events [2]. In other words, probabilities add for mutually exclusive events. A **probability space** is the tuple $(\Omega,\sigma(\Omega),P)$ that encapsulates this structure [2] [3].

# Conditional Probability

The **conditional probability** $P(A\mid B)$ measures the likelihood of event $A$ when we know that event $B$ has occurred. It is defined by $$P(A\mid B) = \frac{P(A \cap B)}{P(B)},$$ provided $P(B)>0$ [4]. Intuitively, it restricts attention to the subset of outcomes where $B$ happens, and asks how often $A$ also happens there. From this definition follows the **product rule** $P(A\cap B)=P(A\mid B)P(B)$ [5], which in turn generalizes to chains of conditioning (e.g. $P(A,B,C)=P(A\mid B,C)P(B\mid C)P(C)$). Conditional probability is fundamental in updating probabilities and in defining *conditional distributions* of random variables.

# Independence

Two events $A$ and $B$ are **independent** if knowing that one occurred does not change the probability of the other. Formally, $A$ and $B$ are independent when $$P(A\cap B) = P(A)\,P(B).$$ Equivalently, $P(A\mid B)=P(A)$ and $P(B\mid A)=P(B)$, meaning $P(A)$ is unaffected by conditioning on $B$ [6]. For example, in two fair dice rolls, the event “first die is 4” is independent of “second die is 3,” so $P(4\;\text{on 1st die and }3\;\text{on 2nd})=P(4\;\text{on 1st})P(3\;\text{on 2nd})$. Checking independence often uses $P(A\cap B)=P(A)P(B)$. Independence greatly simplifies probability computations: independent variables factorize their joint distributions.

# Law of Total Probability

The **law of total probability** breaks a probability into contributions from a partition of the sample space. If events $B_1,\ldots,B_n$ are disjoint and cover $\Omega$, then for any event $A$ we have $$P(A) \;=\; \sum_{i=1}^n P(A\cap B_i) \;=\; \sum_{i=1}^n P(A\mid B_i)\,P(B_i).$$ This formula allows computing $P(A)$ by conditioning on each $B_i$ [7]. For example, if there are several mutually exclusive scenarios (like which die was rolled), the total probability of an outcome is the sum of probabilities under each scenario weighted by the scenario’s chance. The law of total probability is also used to derive marginal probabilities from joint distributions (see *Marginal Probability* below).

# Bayes' Rule

**Bayes' rule** inverts conditional probabilities: for events $A$ and $B$ with $P(B)>0$,
$$P(A\mid B) = \frac{P(B\mid A)\,P(A)}{P(B)}.$$
This follows from $P(A\cap B)=P(A\mid B)P(B)=P(B\mid A)P(A)$ [8]. Bayes' rule lets us update our belief in hypothesis $A$ after observing evidence $B$. In statistical terms, if ${H_i}$ are hypotheses and $D$ is data, Bayes' rule is often written
$$P(H_i\mid D) = \frac{P(D\mid H_i)\,P(H_i)}{\sum_j P(D\mid H_j)\,P(H_j)},$$ where $P(H_i)$ is the *prior* probability of hypothesis $H_i$, $P(D\mid H_i)$ is the *likelihood* of data under $H_i$, and $P(H_i\mid D)$ is the *posterior* probability [9]. This shows how to combine prior beliefs with new data. For instance, if you have two models for data, Bayes' rule gives the updated probability of each model given the observed data, favoring models that make the data more likely (higher likelihood).

# Random Variables

A **random variable** is a function that assigns a numerical value to each outcome in the sample space. Formally, $X:\Omega\to \Lambda$ maps an outcome $\omega$ to a value $X(\omega)$ in some set $\Lambda$. For example, $X$ might count the number of heads in a sequence of coin flips. The random variable induces a probability measure on its values: for any set $A\subseteq \Lambda$, we define $P_X(A)=P({\omega:X(\omega)\in A})$. In particular, if $X$ is *discrete* (taking values in a finite or countable set $\Lambda$), we have a **probability mass function** (PMF) $P_X(x)=P(X=x)$ [10]. For example, if $X$ is the number of heads in three fair coin tosses, then $P_X(2)=3/8$ because there are three outcomes with exactly two heads [11]. Defining random variables allows us to express probabilities and expectations more conveniently than working directly with sample-space events.

# Discrete vs Continuous Random Variables

Random variables come in two main types. A **discrete** random variable takes values in a countable set (like ${0,1,2,\ldots}$), and is described by a PMF $P_X(x)$ giving the probability of each value. In contrast, a **continuous** random variable takes values in an interval (like $\mathbb{R}$), and is described by a **probability density function** (PDF) $f_X(x)$ (if it has one) such that probabilities of intervals are given by integrals of $f_X$. Specifically, for a continuous $X$, the **cumulative distribution function** (CDF) $F_X(x)=P(X\le x)$ is differentiable and its derivative is the PDF: $f_X(x)=F_X'(x)$ [12]. For example, a fair die roll gives a discrete uniform distribution on ${1,\dots,6}$ with $P_X(k)=1/6$, whereas measuring height in a population might yield a continuous Gaussian PDF. The distinction matters because sums become integrals for continuous variables (e.g. $P(a<X<b)=\int_a^b f_X(x)\,dx$), whereas for discrete ones sums of PMF values yield probabilities.

# Joint, Marginal, and Conditional Distributions

For multiple random variables (say $X$ and $Y$), the **joint distribution** describes probabilities of all combinations of values. For discrete variables, the joint probability mass function is $P(X=x,Y=y)$. From a joint distribution one obtains **marginal distributions** by summing (or integrating) out the other variables. For instance, the marginal of $X$ is $$P(X=x)=\sum_{y} P(X=x,Y=y),$$ which "drops" $Y$ as a nuisance variable [13]. This use of the law of total probability is called *marginalization*. The **conditional distribution** of $X$ given $Y=y$ is $P(X=x\mid Y=y)=P(X=x,Y=y)/P(Y=y)$ (for $P(Y=y)>0$). Joint, marginal, and conditional probabilities are related: $P(X,Y)=P(X\mid Y)P(Y)=P(Y\mid X)P(X)$. These concepts

generalize the single-variable definitions. For example, if $X$ is die roll and $Y$ indicates parity (even/odd), one can check independence or compute marginals by summing over $Y$ or $X$ [14] [15].

# Expectation, Variance, and Covariance

The **expectation** or **mean** of a random variable $X$ is its long-run average value. For discrete $X$, it is $$E[X]=\sum_x x\,P(X=x).$$ For example, a fair die has $E[X]=3.5$. Expectation measures central tendency. The **variance** measures spread: $\mathrm{Var}[X]=E[(X-E[X])^2]=E[X^2]-E[X]^2$ [16]. It is the average squared deviation from the mean. The **standard deviation** is the square-root of variance. In the three-coin example above, $X$ (number of heads) has $E[X]=1.5$ and $\mathrm{Var}(X)=0.75$ [17] [18]. More generally, moments are expectations of powers of $X$. The **covariance** between two variables $X$ and $Y$ quantifies their joint variability: $$\mathrm{Cov}[X,Y]=E[(X-E[X])(Y-E[Y])] = E[XY] - E[X]E[Y].$$ By symmetry $\mathrm{Var}[X]=\mathrm{Cov}[X,X]$. Covariance is positive if $X,Y$ tend to move together and negative if oppositely. For jointly distributed $X,Y$, covariance appears in the formula for the variance of a sum: $\mathrm{Var}[X+Y]=\mathrm{Var}[X]+\mathrm{Var}[Y]+2\mathrm{Cov}[X,Y]$.

# Bernoulli Distribution

The **Bernoulli distribution** models a single yes/no trial (like a coin flip). A random variable $X\in\{0,1\}$ is Bernoulli with parameter $p$ if
$$P(X=1)=p,\quad P(X=0)=1-p,$$
often written $P_X(x)=p^x(1-p)^{1-x}$ for $x=0,1$ [19]. Here $p$ (with $0\le p\le1$) is the probability of "success" ($X=1$). If the coin is fair, $p=1/2$. The Bernoulli has mean $E[X]=p$ and variance $p(1-p)$ [20]. It is the simplest discrete distribution and forms the basis for binomial and categorical models.

# Beta Distribution

The **Beta distribution** is a family of continuous distributions on $[0,1]$, often used to model random probabilities. Its probability density is
$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}, \quad 0\le x\le1,$$
where $\alpha,\beta>0$ are shape parameters [21]. Intuitively, $\alpha-1$ and $\beta-1$ can be thought of as "prior counts" of successes and failures. The Beta is a conjugate prior for Bernoulli trials. Its mean is $\frac{\alpha}{\alpha+\beta}$ and variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ { [22]. For example, $\mathrm{Beta}(1,1)$ is uniform on $[0,1]$, while larger parameters concentrate the density around the mean.

# Categorical (Multinoulli) Distribution

The **Categorical** (or *multinoulli*) distribution generalizes Bernoulli to more than two outcomes. A random variable $X$ taking values in ${1,\dots,n}$ follows a categorical distribution with parameters $p_1,\dots,p_n$ (summing to 1) if
$$P(X=i)=p_i,\quad i=1,\dots,n.$$ Equivalently, $P_X(x)=p_x$ for each category $x$ [23]. The parameters $p_i$ represent the probability of each outcome (like sides of a loaded die). The expected value of $X$ is $\sum_i i\,p_i$, and the variance is $\sum_i i^2 p_i -(\sum_i i p_i)^2$ [24]. When $n=2$, this reduces to the Bernoulli case.

# Dirichlet Distribution

The **Dirichlet distribution** is a continuous distribution over probability vectors. A random vector $X=(X_1,\dots,X_d)$ is Dirichlet with parameters $\alpha_1,\dots,\alpha_d>0$ if its density on the simplex ($X_i\ge0$, $\sum X_i=1$) is
$$f_X(x) \;=\; \frac{\Gamma(\sum_i\alpha_i)}{\prod_{i=1}^d\Gamma(\alpha_i)}\,\prod_{i=1}^d x_i^{\alpha_i-1},$$
for $x_i\ge0$, $\sum x_i=1$ [25]. Intuitively, it is a “multi-coin” generalization of the Beta. Each $\alpha_i-1$ acts like a pseudo-count of category $i$. The Dirichlet is the conjugate prior for the categorical distribution. Its mean is $E[X_i]=\alpha_i/(\sum_j\alpha_j)$ [26], giving the expected proportions. Larger $\alpha$ values make the probability vector more concentrated around the mean.

# Uniform Distribution

The **continuous uniform distribution** on an interval $[a,b]$ assigns equal probability density to every point in $[a,b]$. Its PDF is
$$f_X(x) = \frac{1}{b-a},\quad a\le x\le b,$$
and $0$ otherwise [27]. All subintervals of the same length are equally likely. The mean is $(a+b)/2$ and variance $(b-a)^2/12$ [28]. This models situations of complete symmetry, e.g. a random pick from $[a,b]$ with no bias.

# Gaussian (Normal) Distribution

The **Gaussian** or **normal distribution** is a bell-shaped continuous distribution on $\mathbb{R}$. Its PDF is
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp!\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big),$$
where $\mu$ is the mean and $\sigma^2>0$ the variance [29]. The curve is symmetric around $\mu$; about 68% of its mass lies within one standard deviation ($\sigma$) of $\mu$. The Gaussian is widely used because of the central limit theorem, which says sums of many independent effects tend to look normal [30]. Its mean and variance parameters determine its center and spread. Many measurement errors and natural phenomena are approximately Gaussian, making it ubiquitous in statistics.

# Multivariate Gaussian Distribution

The **multivariate Gaussian** (or normal) generalizes the normal to $\mathbb{R}^n$. A random vector $X\in\mathbb{R}^n$ has a multivariate normal distribution with mean vector $\mu\in\mathbb{R}^n$ and covariance matrix $\Sigma\succ0$ if its density is
$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp!\Big(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big)$$
for $x\in\mathbb{R}^n$ [31]. Its mean is $E[X]=\mu$ and covariance $Cov[X]=\Sigma$. A key property is that any marginal or conditional subset of a multivariate normal is also normal [32] [33]. For example, if $(X,Y)$ is jointly normal, then $X$ alone is normal and $Y$ alone is normal. Moreover, the conditional distribution of $X$ given $Y$ is normal, with a mean and covariance adjusted by the observed $Y$ value [33] [34]. This makes the Gaussian convenient for modeling vector-valued measurements and for linear models, because projections and conditionals remain Gaussian.

**Sources:** Definitions and formulae are drawn from standard probability theory references and the provided probability theory notebook [1] [2] [10] [35] [19] [29] [31], among others, adapted here for clarity.

---

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] 04 Probability Theory.ipynb
file://file-1XY8kP5sSvquY46X81fMQA