1 Probability Distributions : Summary

• *Discrete distributions*: Let n label the distinct possible outcomes of a discrete random process, and let p_n be the probability for outcome n. Let A be a quantity which takes values which depend on n, with A_n being the value of A under the outcome n. Then the expected value of A is $\langle A \rangle = \sum_n p_n A_n$, where the sum is over all possible allowed values of n. We must have that the distribution is normalized, *i.e.* $\langle 1 \rangle = \sum_n p_n = 1$.

• *Continuous distributions*: When the random variable $\boldsymbol{\varphi}$ takes a continuum of values, we define the *probability density* $P(\boldsymbol{\varphi})$ to be such that $P(\boldsymbol{\varphi})\, d\mu$ is the probability for the outcome to lie within a differential volume $d\mu$ of $\boldsymbol{\varphi}$, where $d\mu = W(\boldsymbol{\varphi}) \prod_{i=1}^{n} d\varphi_i$, were $\boldsymbol{\varphi}$ is an n-component vector in the configuration space Ω, and where the function $W(\boldsymbol{\varphi})$ accounts for the possibility of different configuration space measures. Then if $A(\boldsymbol{\varphi})$ is any function on Ω, the expected value of A is $\langle A \rangle = \int_\Omega d\mu\, P(\boldsymbol{\varphi})\, A(\boldsymbol{\varphi})$.

• *Central limit theorem*: If $\{x_1, \ldots, x_N\}$ are each independently distributed according to $P(x)$, then the distribution of the sum $X = \sum_{i=1}^N x_i$ is

$$\mathcal{P}_N(X) = \int_{-\infty}^{\infty} \! dx_1 \cdots \int_{-\infty}^{\infty} \! dx_N \, P(x_1) \cdots P(x_N)\, \delta\Big(X - \sum_{i=1}^N x_i\Big) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} \, \Big[\hat{P}(k)\Big]^N e^{ikX} \quad ,$$

where $\hat{P}(k) = \int dx \, P(x)\, e^{-ikx}$ is the Fourier transform of $P(x)$. Assuming that the lowest moments of $P(x)$ exist, $\ln\big[\hat{P}(k)\big] = -i\mu k - \frac{1}{2}\sigma^2 k^2 + \mathcal{O}(k^3)$, where $\mu = \langle x \rangle$ and $\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2$ are the mean and standard deviation. Then for $N \to \infty$,

$$P_N(X) = (2\pi N \sigma^2)^{-1/2}\, e^{-(X - N\mu)^2/2N\sigma^2} \quad ,$$

which is a Gaussian with mean $\langle X \rangle = N\mu$ and standard deviation $\sqrt{\langle X^2 \rangle - \langle X \rangle^2} = \sqrt{N}\,\sigma$. Thus, X is distributed as a Gaussian, even if $P(x)$ is not a Gaussian itself.

• *Entropy*: The entropy of a statistical distribution is $\{p_n\}$ is $S = -\sum_n p_n \ln p_n$. (Sometimes the base 2 logarithm is used, in which case the entropy is measured in *bits*.) This has the interpretation of the *information content* per element of a random sequence.

• *Distributions from maximum entropy*: Given a distribution $\{p_n\}$ subject to $(K+1)$ constraints of the form $\mathcal{X}^a = \sum_n X_n^a\, p_n$ with $a \in \{0, \ldots, K\}$, where $\mathcal{X}^0 = X_n^0 = 1$ (normalization), the distribution consistent with these constraints which maximizes the entropy function is obtained by extremizing the multivariable function

$$S^*\big(\{p_n\}, \{\lambda_a\}\big) = -\sum_n p_n \ln p_n - \sum_{a=0}^K \lambda_a \Big(\sum_n X_n^a\, p_n - \mathcal{X}^a \Big) \quad ,$$

with respect to the probabilities $\{p_n\}$ and the Lagrange multipliers $\{\lambda_a\}$. This results in a Gibbs distribution,

$$p_n = \frac{1}{Z} \exp\Big\{ -\sum_{a=1}^K \lambda_a X_n^a \Big\} \quad ,$$

where $Z = e^{1+\lambda_0}$ is determined by normalization, *i.e.* $\sum_n p_n = 1$ (*i.e.* the $a = 0$ constraint) and the K remaining multipliers determined by the K additional constraints.

• *Multidimensional Gaussian integral*:

$$\int\limits_{-\infty}^{\infty} dx_1 \cdots \int\limits_{-\infty}^{\infty} dx_n \, \exp\Big(-\tfrac{1}{2}\, x_i\, A_{ij}\, x_j + b_i\, x_i\Big) = \left(\frac{(2\pi)^n}{\det A}\right)^{1/2} \exp\Big(\tfrac{1}{2}\, b_i\, A^{-1}_{ij}\, b_j\Big) \quad .$$

• *Bayes' theorem*: Let the *conditional probability* for B given A be $P(B|A)$. Then Bayes' theorem says $P(A|B) = P(A) \cdot P(B|A)\, /\, P(B)$. If the 'event space' is partitioned as $\{A_i\}$, then we have the extended form,

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_j P(B|A_j) \cdot P(A_j)} \quad .$$

When the event space is a 'binary partition' $\{A, \neg A\}$, as is often the case in fields like epidemiology (*i.e.* test positive or test negative), we have

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)} \quad .$$

Note that $P(A|B) + P(\neg A|B) = 1$ (which follows from $\neg\neg A = A$).

• *Updating Bayesian priors*: Given data in the form of observed values $\boldsymbol{x} = \{x_1, \ldots, x_N\} \in \mathcal{X}$ and a hypothesis in the form of parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_K\} \in \Theta$, we write the conditional probability (density) for observing $\boldsymbol{x}$ given $\boldsymbol{\theta}$ as $f(\boldsymbol{x}|\boldsymbol{\theta})$. Bayes' theorem says that the corresponding distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ for $\boldsymbol{\theta}$ conditioned on $\boldsymbol{x}$ is

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})}{\int\limits_{\Theta} d\boldsymbol{\theta}'\, f(\boldsymbol{x}|\boldsymbol{\theta}')\, \pi(\boldsymbol{\theta}')} \quad ,$$

We call $\pi(\boldsymbol{\theta})$ the *prior* for $\boldsymbol{\theta}$, $f(\boldsymbol{x}|\boldsymbol{\theta})$ the *likelihood* of $\boldsymbol{x}$ given $\boldsymbol{\theta}$, and $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ the *posterior* for $\boldsymbol{\theta}$ given $\boldsymbol{x}$. We can use the posterior to find the distribution of new data points $\boldsymbol{y}$, called the *posterior predictive distribution*, $f(\boldsymbol{y}|\boldsymbol{x}) = \int\limits_{\Theta} d\boldsymbol{\theta}\, f(\boldsymbol{y}|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta}|\boldsymbol{x})$. This is the update of the *prior predictive distribution*, $f(\boldsymbol{x}) = \int\limits_{\Theta} d\boldsymbol{\theta}\, f(\boldsymbol{x}|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$. As an example, consider coin flipping with $f(\boldsymbol{x}|\boldsymbol{\theta}) = \theta^X\, (1-\theta)^{N-X}$, where N is the number of flips, and $X = \sum_{j=1}^{N} x_j$ with x_j a discrete variable which is 0 for tails and 1 for heads. The parameter $\theta \in [0, 1]$ is the probability to flip heads. We choose a prior $\pi(\theta) = \theta^{\alpha-1}\, (1-\theta)^{\beta-1} / \mathsf{B}(\alpha, \beta)$ where $\mathsf{B}(\alpha, \beta) = \Gamma(\alpha)\, \Gamma(\beta) / \Gamma(\alpha+\beta)$ is the Beta distribution. This results in a normalized prior $\int\limits_0^1 d\theta\, \pi(\theta) = 1$. The posterior distribution for θ is then

$$\pi(\theta|x_1, \ldots, x_N) = \frac{f(x_1, \ldots, x_N|\theta)\, \pi(\theta)}{\int_0^1 d\theta'\, f(x_1, \ldots, x_N|\theta')\, \pi(\theta')} = \frac{\theta^{X+\alpha-1}(1-\theta)^{N-X+\beta-1}}{\mathsf{B}(X+\alpha, N-X+\beta)} \quad .$$

The prior predictive is $f(\boldsymbol{x}) = \int_0^1 d\theta f(\boldsymbol{x}|\theta)\,\pi(\theta) = \mathsf{B}(X+\alpha, N-X+\beta)/\mathsf{B}(\alpha,\beta)$, and the posterior predictive for the total number of heads Y in M flips is

$$f(\boldsymbol{y}|\boldsymbol{x}) = \int_0^1 d\theta\, f(\boldsymbol{y}|\theta)\,\pi(\theta|\boldsymbol{x}) = \frac{\mathsf{B}(X+Y+\alpha, N-X+M-Y+\beta)}{\mathsf{B}(X+\alpha, N-X+\beta)} \quad .$$