1 Probability Distributions: Summary

The properties of a statistical distribution depend on

• 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. $\sum_{n} p_{n} = 1$.

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

• 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

\[
P_N(X) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P(x_1) \cdots P(x_N) \delta \left( X - \sum_{i=1}^{N} x_i \right) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} \left[ \hat{P}(k) \right]^{N} e^{ikX},
\]

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

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

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 $\lambda^{a} = \sum_{n} X_{n}^{a} p_{n}$ with $a \in \{0, \ldots, K\}$, where $\lambda^{0} = X_{n}^{0} = 1$ (normalization), the distribution consistent with these constraints which maximizes the entropy function is obtained by extremizing the multivariable function

\[
S^{*}(\{p_{n}\}; \{\lambda_{a}\}) = -\sum_{n} p_{n} \ln p_{n} - \sum_{a=0}^{K} \lambda_{a} \left( \sum_{n} X_{n}^{a} p_{n} - \lambda^{a} \right),
\]

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

\[
p_{n} = \frac{1}{Z} \exp \left\{ -\sum_{a=1}^{K} \lambda_{a} X_{n}^{a} \right\},
\]
where $Z = e^{1+λ}$ 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:**
  $$\int_{-\infty}^{\infty} dx_1 \cdots \int_{-\infty}^{\infty} dx_n \exp\left(-\frac{1}{2} x_i A_{ij} x_j + b_i x_i\right) = \left(\frac{(2\pi)^n}{\text{det}A}\right)^{1/2} \exp\left(\frac{1}{2} b_i A_{ij}^{-1} b_j\right).$$

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

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

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 $x = \{x_1, \ldots, x_N\} \in \mathcal{X}$ and a hypothesis in the form of parameters $\theta = \{θ_1, \ldots, θ_K\} \in \Theta$, we write the conditional probability (density) for observing $x$ given $θ$ as $f(x|θ)$. Bayes’ theorem says that the corresponding distribution $π(θ|x)$ for $θ$ conditioned on $x$ is
  $$π(θ|x) = \frac{f(x|θ) \cdot π(θ)}{\int f(x|θ') \cdot π(θ') \, dθ'}.$$  

We call $π(θ)$ the prior for $θ$, $f(x|θ)$ the likelihood of $x$ given $θ$, and $π(θ|x)$ the posterior for $θ$ given $x$. We can use the posterior to find the distribution of new data points $y$, called the posterior predictive distribution, $f(y|x) = \int f(θ|y) \cdot π(θ|x) \, dθ$. This is the update of the prior predictive distribution, $f(x) = \int f(θ) \cdot π(θ) \, dθ$. As an example, consider coin flipping with $f(x|θ) = θ^X (1-θ)^{N-X}$, 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 $θ \in [0, 1]$ is the probability to flip heads. We choose a prior $π(θ) = θ^{α-1} (1-θ)^{β-1} / B(α, β)$ where $B(α, β) = Γ(α) \Gamma(β)/Γ(α+β)$ is the Beta distribution. This results in a normalized prior $\int_0^1 dθ \cdot π(θ) = 1$. The posterior distribution for $θ$ is then
  $$π(θ|x_1, \ldots, x_N) = \frac{f(x_1, \ldots, x_N|θ) \cdot π(θ)}{\int_0^1 dθ' \cdot f(x_1, \ldots, x_N|θ') \cdot π(θ')} = \frac{θ^{X+α-1} (1-θ)^{N-X+β-1}}{B(X+α, N-X+β)}.$$  

2
The prior predictive is \( f(x) = \frac{1}{\theta} \int_{0}^{\infty} d\theta f(x|\theta) \pi(\theta) = B(X + \alpha, N - X + \beta)/B(\alpha, \beta) \), and the posterior predictive for the total number of heads \( Y \) in \( M \) flips is

\[
f(y|x) = \int_{0}^{1} d\theta f(y|\theta) \pi(\theta|x) = \frac{B(X + Y + \alpha, N - X + M - Y + \beta)}{B(X + \alpha, N - X + \beta)} .
\]