Probability Distributions

In this appendix, we summarize the main properties of some of the most widely used probability distributions, and for each distribution we list some key statistics such as the expectation $\mathbb{E}[\mathbf{x}]$, the variance (or covariance), the mode, and the entropy $\mathrm{H}[\mathbf{x}]$. All of these distributions are members of the exponential family and are widely used as building blocks for more sophisticated probabilistic models.

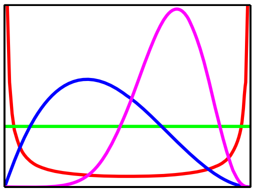
Bernoulli

This is the distribution for a single binary variable $x \in \{0,1\}$ representing, for example, the result of flipping a coin. It is governed by a single continuous parameter $\mu \in [0,1]$ that represents the probability of $x = 1$.

$$\begin{aligned}
\mathrm{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} && \text{(B.1)}\\
\mathbb{E}[x] &= \mu && \text{(B.2)}\\
\mathrm{var}[x] &= \mu(1-\mu) && \text{(B.3)}\\
\mathrm{mode}[x] &= \begin{cases} 1 & \text{if } \mu \geqslant 0.5, \\ 0 & \text{otherwise} \end{cases} && \text{(B.4)}\\
\mathrm{H}[x] &= -\mu\ln\mu - (1-\mu)\ln(1-\mu). && \text{(B.5)}
\end{aligned}$$

The Bernoulli is a special case of the binomial distribution for the case of a single observation. Its conjugate prior for μ is the beta distribution.

Beta

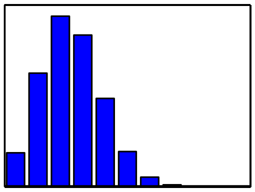


This is a distribution over a continuous variable $\mu \in [0, 1]$, which is often used to represent the probability for some binary event. It is governed by two parameters a and b that are constrained by $a > 0$ and $b > 0$ to ensure that the distribution can be normalized.

$$\begin{aligned} \mathrm{Beta}(\mu|a, b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} && \text{(B.6)} \\ \mathbb{E}[\mu] &= \frac{a}{a+b} && \text{(B.7)} \\ \mathrm{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)} && \text{(B.8)} \\ \mathrm{mode}[\mu] &= \frac{a-1}{a+b-2}. && \text{(B.9)} \end{aligned}$$

The beta is the conjugate prior for the Bernoulli distribution, for which a and b can be interpreted as the effective prior number of observations of $x = 1$ and $x = 0$, respectively. Its density is finite if $a \geqslant 1$ and $b \geqslant 1$, otherwise there is a singularity at $\mu = 0$ and/or $\mu = 1$. For $a = b = 1$, it reduces to a uniform distribution. The beta distribution is a special case of the K-state Dirichlet distribution for $K = 2$.

Binomial



The binomial distribution gives the probability of observing m occurrences of $x = 1$ in a set of N samples from a Bernoulli distribution, where the probability of observing $x = 1$ is $\mu \in [0, 1]$.

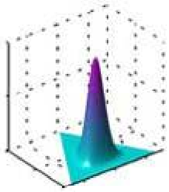
$$\begin{aligned} \mathrm{Bin}(m|N, \mu) &= \binom{N}{m}\mu^m(1-\mu)^{N-m} && \text{(B.10)} \\ \mathbb{E}[m] &= N\mu && \text{(B.11)} \\ \mathrm{var}[m] &= N\mu(1-\mu) && \text{(B.12)} \\ \mathrm{mode}[m] &= \lfloor (N+1)\mu \rfloor && \text{(B.13)} \end{aligned}$$

where $\lfloor (N+1)\mu \rfloor$ denotes the largest integer that is less than or equal to $(N+1)\mu$, and the quantity

$$\binom{N}{m} = \frac{N!}{m!(N-m)!} \qquad \text{(B.14)}$$

denotes the number of ways of choosing m objects out of a total of N identical objects. Here $m!$, pronounced 'factorial m', denotes the product $m \times (m-1) \times \ldots, \times 2 \times 1$. The particular case of the binomial distribution for $N = 1$ is known as the Bernoulli distribution, and for large N the binomial distribution is approximately Gaussian. The conjugate prior for μ is the beta distribution.

Dirichlet



The Dirichlet is a multivariate distribution over K random variables $0 \leqslant \mu_k \leqslant 1$, where $k = 1, \ldots, K$, subject to the constraints

$$0 \leqslant \mu_k \leqslant 1, \qquad \sum_{k=1}^{K} \mu_k = 1. \tag{B.15}$$

Denoting $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^{\mathrm{T}}$, we have

$$\begin{aligned}
\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) &= C(\boldsymbol{\alpha}) \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} && \text{(B.16)}\\
\mathbb{E}[\mu_k] &= \frac{\alpha_k}{\widehat{\alpha}} && \text{(B.17)}\\
\mathrm{var}[\mu_k] &= \frac{\alpha_k(\widehat{\alpha} - \alpha_k)}{\widehat{\alpha}^2(\widehat{\alpha} + 1)} && \text{(B.18)}\\
\mathrm{cov}[\mu_j \mu_k] &= -\frac{\alpha_j \alpha_k}{\widehat{\alpha}^2(\widehat{\alpha} + 1)} && \text{(B.19)}\\
\mathrm{mode}[\mu_k] &= \frac{\alpha_k - 1}{\widehat{\alpha} - K} && \text{(B.20)}\\
\mathbb{E}[\ln \mu_k] &= \psi(\alpha_k) - \psi(\widehat{\alpha}) && \text{(B.21)}\\
\mathrm{H}[\boldsymbol{\mu}] &= -\sum_{k=1}^{K} (\alpha_k - 1)\left\{\psi(\alpha_k) - \psi(\widehat{\alpha})\right\} - \ln C(\boldsymbol{\alpha}) && \text{(B.22)}
\end{aligned}$$

where

$$C(\boldsymbol{\alpha}) = \frac{\Gamma(\widehat{\alpha})}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \tag{B.23}$$

and

$$\widehat{\alpha} = \sum_{k=1}^{K} \alpha_k. \tag{B.24}$$

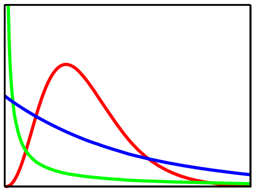
Here

$$\psi(a) \equiv \frac{d}{da} \ln \Gamma(a) \tag{B.25}$$

is known as the *digamma* function (Abramowitz and Stegun, 1965). The parameters α_k are subject to the constraint $\alpha_k > 0$ in order to ensure that the distribution can be normalized.

The Dirichlet forms the conjugate prior for the multinomial distribution and represents a generalization of the beta distribution. In this case, the parameters α_k can be interpreted as effective numbers of observations of the corresponding values of the K-dimensional binary observation vector $\mathbf{x}$. As with the beta distribution, the Dirichlet has finite density everywhere provided $\alpha_k \geqslant 1$ for all k.

Gamma

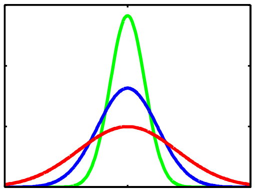


The Gamma is a probability distribution over a positive random variable $\tau > 0$ governed by parameters a and b that are subject to the constraints $a > 0$ and $b > 0$ to ensure that the distribution can be normalized.

$$\begin{aligned} \mathrm{Gam}(\tau|a,b) &= \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau} && \text{(B.26)} \\ \mathbb{E}[\tau] &= \frac{a}{b} && \text{(B.27)} \\ \mathrm{var}[\tau] &= \frac{a}{b^2} && \text{(B.28)} \\ \mathrm{mode}[\tau] &= \frac{a-1}{b} \quad \text{for} \quad \alpha \geqslant 1 && \text{(B.29)} \\ \mathbb{E}[\ln \tau] &= \psi(a) - \ln b && \text{(B.30)} \\ \mathrm{H}[\tau] &= \ln \Gamma(a) - (a-1)\psi(a) - \ln b + a && \text{(B.31)} \end{aligned}$$

where $\psi(\cdot)$ is the digamma function defined by (B.25). The gamma distribution is the conjugate prior for the precision (inverse variance) of a univariate Gaussian. For $a \geqslant 1$ the density is everywhere finite, and the special case of $a = 1$ is known as the *exponential* distribution.

Gaussian



The Gaussian is the most widely used distribution for continuous variables. It is also known as the *normal* distribution. In the case of a single variable $x \in (-\infty, \infty)$ it is governed by two parameters, the mean $\mu \in (-\infty, \infty)$ and the variance $\sigma^2 > 0$.

$$\begin{aligned} \mathcal{N}(x|\mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} && \text{(B.32)} \\ \mathbb{E}[x] &= \mu && \text{(B.33)} \\ \mathrm{var}[x] &= \sigma^2 && \text{(B.34)} \\ \mathrm{mode}[x] &= \mu && \text{(B.35)} \\ \mathrm{H}[x] &= \frac{1}{2} \ln \sigma^2 + \frac{1}{2}\left(1 + \ln(2\pi)\right). && \text{(B.36)} \end{aligned}$$

The inverse of the variance $\tau = 1/\sigma^2$ is called the precision, and the square root of the variance σ is called the standard deviation. The conjugate prior for μ is the Gaussian, and the conjugate prior for τ is the gamma distribution. If both μ and τ are unknown, their joint conjugate prior is the Gaussian-gamma distribution.

For a D-dimensional vector $\mathbf{x}$, the Gaussian is governed by a D-dimensional mean vector $\boldsymbol{\mu}$ and a $D \times D$ covariance matrix $\boldsymbol{\Sigma}$ that must be symmetric and positive-definite.

$$\begin{aligned}
\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} && \text{(B.37)}\\
\mathbb{E}[\mathbf{x}] &= \boldsymbol{\mu} && \text{(B.38)}\\
\operatorname{cov}[\mathbf{x}] &= \boldsymbol{\Sigma} && \text{(B.39)}\\
\operatorname{mode}[\mathbf{x}] &= \boldsymbol{\mu} && \text{(B.40)}\\
\mathrm{H}[\mathbf{x}] &= \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{D}{2}\left(1+\ln(2\pi)\right). && \text{(B.41)}
\end{aligned}$$

The inverse of the covariance matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix, which is also symmetric and positive definite. Averages of random variables tend to a Gaussian, by the central limit theorem, and the sum of two Gaussian variables is again Gaussian. The Gaussian is the distribution that maximizes the entropy for a given variance (or covariance). Any linear transformation of a Gaussian random variable is again Gaussian. The marginal distribution of a multivariate Gaussian with respect to a subset of the variables is itself Gaussian, and similarly the conditional distribution is also Gaussian. The conjugate prior for $\boldsymbol{\mu}$ is the Gaussian, the conjugate prior for $\boldsymbol{\Lambda}$ is the Wishart, and the conjugate prior for $(\boldsymbol{\mu},\boldsymbol{\Lambda})$ is the Gaussian-Wishart.

If we have a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}) && \text{(B.42)}\\
p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b},\mathbf{L}^{-1}) && \text{(B.43)}
\end{aligned}$$

then the marginal distribution of $\mathbf{y}$, and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$, are given by

$$\begin{aligned}
p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b},\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) && \text{(B.44)}\\
p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\},\boldsymbol{\Sigma}) && \text{(B.45)}
\end{aligned}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}. \qquad \text{(B.46)}$$

If we have a joint Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$ and we define the following partitions

$$\mathbf{x} = \begin{pmatrix}\mathbf{x}_a\\ \mathbf{x}_b\end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{pmatrix} \qquad \text{(B.47)}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab}\\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{pmatrix}, \qquad \boldsymbol{\Lambda} = \begin{pmatrix}\boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab}\\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb}\end{pmatrix} \qquad \text{(B.48)}$$

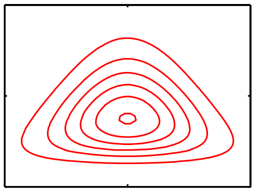
then the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ is given by

$$\begin{aligned}
p(\mathbf{x}_a|\mathbf{x}_b) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b},\boldsymbol{\Lambda}_{aa}^{-1}) && \text{(B.49)}\\
\boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) && \text{(B.50)}
\end{aligned}$$

and the marginal distribution $p(\mathbf{x}_a)$ is given by

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}). \tag{B.51}$$

Gaussian-Gamma



This is the conjugate prior distribution for a univariate Gaussian $\mathcal{N}(x|\mu, \lambda^{-1})$ in which the mean μ and the precision λ are both unknown and is also called the *normal-gamma* distribution. It comprises the product of a Gaussian distribution for μ, whose precision is proportional to λ, and a gamma distribution over λ.

$$p(\mu, \lambda|\mu_0, \beta, a, b) = \mathcal{N}\left(\mu|\mu_o, (\beta\lambda)^{-1}\right) \mathrm{Gam}(\lambda|a, b). \tag{B.52}$$

Gaussian-Wishart

This is the conjugate prior distribution for a multivariate Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda})$ in which both the mean $\boldsymbol{\mu}$ and the precision $\boldsymbol{\Lambda}$ are unknown, and is also called the normal-Wishart distribution. It comprises the product of a Gaussian distribution for $\boldsymbol{\mu}$, whose precision is proportional to $\boldsymbol{\Lambda}$, and a Wishart distribution over $\boldsymbol{\Lambda}$.

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{\mu}_0, \beta, \mathbf{W}, \nu) = \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\beta\boldsymbol{\Lambda})^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu). \tag{B.53}$$

For the particular case of a scalar x, this is equivalent to the Gaussian-gamma distribution.

Multinomial

If we generalize the Bernoulli distribution to an K-dimensional binary variable $\mathbf{x}$ with components $x_k \in \{0, 1\}$ such that $\sum_k x_k = 1$, then we obtain the following discrete distribution

$$\begin{aligned}
p(\mathbf{x}) &= \prod_{k=1}^{K} \mu_k^{x_k} && \text{(B.54)}\\
\mathbb{E}[x_k] &= \mu_k && \text{(B.55)}\\
\mathrm{var}[x_k] &= \mu_k(1-\mu_k) && \text{(B.56)}\\
\mathrm{cov}[x_j x_k] &= I_{jk}\mu_k && \text{(B.57)}\\
\mathrm{H}[\mathbf{x}] &= -\sum_{k=1}^{M} \mu_k \ln \mu_k && \text{(B.58)}
\end{aligned}$$

where I_{jk} is the j, k element of the identity matrix. Because $p(x_k = 1) = \mu_k$, the parameters must satisfy $0 \leqslant \mu_k \leqslant 1$ and $\sum_k \mu_k = 1$.

The multinomial distribution is a multivariate generalization of the binomial and gives the distribution over counts m_k for a K-state discrete variable to be in state k given a total number of observations N.

$$\begin{aligned}
\mathrm{Mult}(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) &= \binom{N}{m_1 m_2 \ldots m_M} \prod_{k=1}^{M} \mu_k^{m_k} && \text{(B.59)}\\
\mathbb{E}[m_k] &= N\mu_k && \text{(B.60)}\\
\mathrm{var}[m_k] &= N\mu_k(1-\mu_k) && \text{(B.61)}\\
\mathrm{cov}[m_j m_k] &= -N\mu_j\mu_k && \text{(B.62)}
\end{aligned}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}$, and the quantity

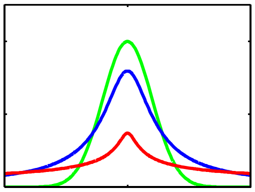
$$\binom{N}{m_1 m_2 \ldots m_K} = \frac{N!}{m_1! \ldots m_K!} \qquad \text{(B.63)}$$

gives the number of ways of taking N identical objects and assigning m_k of them to bin k for $k = 1, \ldots, K$. The value of μ_k gives the probability of the random variable taking state k, and so these parameters are subject to the constraints $0 \leqslant \mu_k \leqslant 1$ and $\sum_k \mu_k = 1$. The conjugate prior distribution for the parameters $\{\mu_k\}$ is the Dirichlet.

Normal

The normal distribution is simply another name for the Gaussian. In this book, we use the term Gaussian throughout, although we retain the conventional use of the symbol $\mathcal{N}$ to denote this distribution. For consistency, we shall refer to the normal-gamma distribution as the Gaussian-gamma distribution, and similarly the normal-Wishart is called the Gaussian-Wishart.

Student's t



This distribution was published by William Gosset in 1908, but his employer, Guiness Breweries, required him to publish under a pseudonym, so he chose 'Student'. In the univariate form, Student's t-distribution is obtained by placing a conjugate gamma prior over the precision of a univariate Gaussian distribution and then integrating out the precision variable. It can therefore be viewed as an infinite mixture of Gaussians having the same mean but different variances.

$$\begin{aligned}
\mathrm{St}(x|\mu,\lambda,\nu) &= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2} && \text{(B.64)}\\
\mathbb{E}[x] &= \mu \quad \text{for} \;\; \nu > 1 && \text{(B.65)}\\
\mathrm{var}[x] &= \frac{1}{\lambda}\frac{\nu}{\nu-2} \quad \text{for} \;\; \nu > 2 && \text{(B.66)}\\
\mathrm{mode}[x] &= \mu. && \text{(B.67)}
\end{aligned}$$

Here $\nu > 0$ is called the number of degrees of freedom of the distribution. The particular case of $\nu = 1$ is called the *Cauchy* distribution.

For a D-dimensional variable $\mathbf{x}$, Student's t-distribution corresponds to marginalizing the precision matrix of a multivariate Gaussian with respect to a conjugate Wishart prior and takes the form

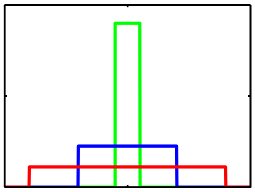
$$\begin{aligned}
\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) &= \frac{\Gamma(\nu/2+D/2)}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(\nu\pi)^{D/2}}\left[1+\frac{\Delta^2}{\nu}\right]^{-\nu/2-D/2} && \text{(B.68)}\\
\mathbb{E}[\mathbf{x}] &= \boldsymbol{\mu} \quad \text{for} \;\; \nu > 1 && \text{(B.69)}\\
\mathrm{cov}[\mathbf{x}] &= \frac{\nu}{\nu-2}\boldsymbol{\Lambda}^{-1} \quad \text{for} \;\; \nu > 2 && \text{(B.70)}\\
\mathrm{mode}[\mathbf{x}] &= \boldsymbol{\mu} && \text{(B.71)}
\end{aligned}$$

where Δ^2 is the squared Mahalanobis distance defined by

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}). \qquad \text{(B.72)}$$

In the limit $\nu \to \infty$, the t-distribution reduces to a Gaussian with mean μ and precision $\boldsymbol{\Lambda}$. Student's t-distribution provides a generalization of the Gaussian whose maximum likelihood parameter values are robust to outliers.

Uniform

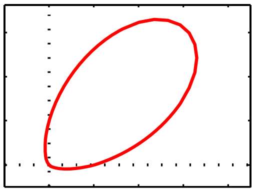


This is a simple distribution for a continuous variable x defined over a finite interval $x \in [a, b]$ where $b > a$.

$$\begin{aligned}
\mathrm{U}(x|a,b) &= \frac{1}{b-a} && \text{(B.73)}\\
\mathbb{E}[x] &= \frac{(b+a)}{2} && \text{(B.74)}\\
\mathrm{var}[x] &= \frac{(b-a)^2}{12} && \text{(B.75)}\\
\mathrm{H}[x] &= \ln(b-a). && \text{(B.76)}
\end{aligned}$$

If x has distribution $\mathrm{U}(x|0,1)$, then $a + (b-a)x$ will have distribution $\mathrm{U}(x|a,b)$.

Von Mises



The von Mises distribution, also known as the circular normal or the circular Gaussian, is a univariate Gaussian-like periodic distribution for a variable $\theta \in [0, 2\pi)$.

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left\{m \cos(\theta - \theta_0)\right\} \qquad \text{(B.77)}$$

where $I_0(m)$ is the zeroth-order Bessel function of the first kind. The distribution has period 2π so that $p(\theta + 2\pi) = p(\theta)$ for all θ. Care must be taken in interpreting this distribution because simple expectations will be dependent on the (arbitrary) choice of origin for the variable θ. The parameter θ_0 is analogous to the mean of a univariate Gaussian, and the parameter $m > 0$, known as the *concentration* parameter, is analogous to the precision (inverse variance). For large m, the von Mises distribution is approximately a Gaussian centred on θ_0.

Wishart

The Wishart distribution is the conjugate prior for the precision matrix of a multivariate Gaussian.

$$\mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, \nu) = B(\mathbf{W}, \nu)|\mathbf{\Lambda}|^{(\nu-D-1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\mathbf{\Lambda})\right) \qquad \text{(B.78)}$$

where

$$B(\mathbf{W}, \nu) \equiv |\mathbf{W}|^{-\nu/2} \left(2^{\nu D/2} \pi^{D(D-1)/4} \prod_{i=1}^{D} \Gamma\left(\frac{\nu + 1 - i}{2}\right)\right)^{-1} \qquad \text{(B.79)}$$

$$\mathbb{E}[\mathbf{\Lambda}] = \nu\mathbf{W} \qquad \text{(B.80)}$$

$$\mathbb{E}\left[\ln|\mathbf{\Lambda}|\right] = \sum_{i=1}^{D} \psi\left(\frac{\nu + 1 - i}{2}\right) + D\ln 2 + \ln|\mathbf{W}| \qquad \text{(B.81)}$$

$$\mathrm{H}[\mathbf{\Lambda}] = -\ln B(\mathbf{W}, \nu) - \frac{(\nu - D - 1)}{2}\mathbb{E}\left[\ln|\mathbf{\Lambda}|\right] + \frac{\nu D}{2} \qquad \text{(B.82)}$$

where $\mathbf{W}$ is a $D \times D$ symmetric, positive definite matrix, and $\psi(\cdot)$ is the digamma function defined by (B.25). The parameter ν is called the *number of degrees of freedom* of the distribution and is restricted to $\nu > D - 1$ to ensure that the Gamma function in the normalization factor is well-defined. In one dimension, the Wishart reduces to the gamma distribution $\mathrm{Gam}(\lambda|a, b)$ given by (B.26) with parameters $a = \nu/2$ and $b = 1/2W$.