

Probability Distributions

Seungjin Choi

Department of Computer Science
 Pohang University of Science and Technology, Korea
seungjin@postech.ac.kr

- ▶ Summarize the main properties of some of the widely-used probability distributions.
- ▶ Taken from Appendix B in [Bishop's PRML](#).

1 / 25

Bernoulli Distribution

- ▶ Distribution for a [single binary random variable](#) $x \in \{0, 1\}$.
- ▶ A special case of the binomial distribution for the case of a single observation.
- ▶ Governed by a single continuous parameter $\mu \in [0, 1]$ that represents the probability of $x = 1$:

$$\text{Bern}(x|\mu) = \mu^x(1 - \mu)^{1-x}.$$

- ▶ [Conjugate prior](#) for μ is the [beta distribution](#).
- ▶ Key statistics are given by

$$\begin{aligned}\mathbb{E}\{x\} &= \mu, \\ \text{var}\{x\} &= \mu(1 - \mu), \\ H(x) &= -\mu \log \mu - (1 - \mu) \log[1 - \mu].\end{aligned}$$

3 / 25

2 / 25

Bernoulli Distribution (Cont'd)

$$\begin{aligned}\mathbb{E}\{x\} &= \sum_x x p(x) \\ &= 0 \cdot p(x = 0) + 1 \cdot p(x = 1) \\ &= \mu.\end{aligned}$$

$$\begin{aligned}\text{var}(x) &= \mathbb{E}\{x^2\} - \mu^2 \\ &= 0 \cdot p(x = 0) + 1 \cdot p(x = 1) - \mu^2 \\ &= \mu(1 - \mu).\end{aligned}$$

$$\begin{aligned}H(x) &= \mathbb{E}\{-\log p(x)\} \\ &= -p(x = 0) \log p(x = 0) - p(x = 1) \log p(x = 1) \\ &= -\mu \log \mu - (1 - \mu) \log(1 - \mu).\end{aligned}$$

4 / 25

Binomial Distribution

- ▶ 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 $\mu \in [0, 1]$:

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}, \quad \binom{N}{m} = \frac{N!}{m!(N-m)!}.$$

- ▶ Key statistics are given by

$$\begin{aligned} \mathbb{E}\{m\} &= N\mu, \\ \text{var}\{m\} &= N\mu(1 - \mu). \end{aligned}$$

Multinomial Distribution

- ▶ A multivariate generalization of binomial distribution, which gives the distribution over counts m_j for a K -state discrete variable to be in state i given a total number of observations N :

$$\text{Mult}(m_1, m_2, \dots, m_K | \mu, N) = \binom{N}{m_1 m_2 \dots m_K} \prod_{j=1}^K \mu_j^{m_j},$$

where

$$\binom{N}{m_1 m_2 \dots m_K} = \frac{N!}{m_1! m_2! \dots m_K!}.$$

- ▶ Key statistics are given by

$$\begin{aligned} \mathbb{E}\{m_j\} &= N\mu_j, \\ \text{var}\{m_j\} &= N\mu_j(1 - \mu_j), \\ \text{cov}\{m_i, m_j\} &= -N\mu_i\mu_j. \end{aligned}$$

Generalization of Bernoulli Distribution

- ▶ We generalize the Bernoulli distribution to an K -dimensional binary random vector \mathbf{x} with components $x_j \in \{0, 1\}$ such that $\sum_j x_j = 1$:

$$p(\mathbf{x}) = \prod_{j=1}^K \mu_j^{x_j}.$$

- ▶ Key statistics are given by

$$\begin{aligned} \mathbb{E}\{x_j\} &= \mu_j, \\ \text{var}\{x_j\} &= \mu_j(1 - \mu_j), \\ \text{cov}\{x_i, x_j\} &= \delta_{ij}\mu_j, \\ H(\mathbf{x}) &= -\sum_{j=1}^K \mu_j \log \mu_j. \end{aligned}$$

5 / 25

6 / 25

Beta Distribution

- ▶ A distribution over a continuous variable $\mu \in [0, 1]$, which is often used to represent the probability for some binary event.
- ▶ Governed by two parameters $a > 0$ and $b > 0$:

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1 - \mu)^{b-1}.$$

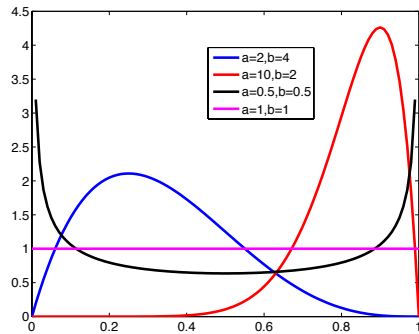
- ▶ Key statistics are given by

$$\begin{aligned} \mathbb{E}\{\mu\} &= \frac{a}{a+b}, \\ \text{var}\{\mu\} &= \frac{ab}{(a+b)^2(a+b+1)}. \end{aligned}$$

7 / 25

8 / 25

Beta Distribution (Cont'd)



- ▶ Density is finite if $a \geq 1$ and $b \geq 1$, otherwise there is a singularity at $\mu = 0$ and/or $\mu = 1$.
- ▶ Reduces to a **uniform distribution** for $a = b = 1$.
- ▶ The **beta distribution** 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.

Beta and Gamma Functions

- ▶ **Beta function** is defined by

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 t^{a-1}(1-t)^{b-1} dt.$$

- ▶ **Gamma function** is defined by

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

which satisfies $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(n+1) = n!$ for natural numbers n .

9 / 25

10 / 25

Compute the mean in the beta

$$\begin{aligned} \mathbb{E}\{\mu\} &= \int_0^1 \mu p(\mu) d\mu \\ &= \frac{1}{B(a, b)} \int_0^1 \mu^a (1-\mu)^{b-1} d\mu \\ &= \frac{B(a+1, b)}{B(a, b)} \\ &= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\ &= \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\ &= \frac{a}{a+b}. \end{aligned}$$

Compute $\mathbb{E}\{\mu^2\}$ in the beta

$$\begin{aligned} \mathbb{E}\{\mu^2\} &= \int_0^1 \mu^2 p(\mu) d\mu \\ &= \frac{1}{B(a, b)} \int_0^1 \mu^{a+1} (1-\mu)^{b-1} d\mu \\ &= \frac{B(a+2, b)}{B(a, b)} \\ &= \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\ &= \frac{(a+1)a\Gamma(a)\Gamma(b)}{(a+b+1)(a+b)\Gamma(a+b)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\ &= \frac{a(a+1)}{(a+b+1)(a+b)}. \end{aligned}$$

11 / 25

12 / 25

Dirichlet Distribution

- ▶ A multivariate distribution over K random variables $0 \leq \mu_j \leq 1$, where $j = 1, \dots, K$, subject to **sum-to-one** constraint $\sum_{j=1}^K \mu_j = 1$.
- ▶ Governed by K parameters $\alpha = [\alpha_1, \dots, \alpha_K]^T$ ($\alpha_j > 0$ for $j = 1, \dots, K$):

$$\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{1}{B(\alpha_1, \dots, \alpha_K)} \prod_{j=1}^K \mu_j^{\alpha_j-1},$$

where

$$B(\alpha_1, \dots, \alpha_K) = B(\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)}{\Gamma(\alpha_1 + \cdots + \alpha_K)}.$$

Dirichlet Distribution (Cont'd)

Key statistics are given by

$$\begin{aligned}\mathbb{E}\{\mu_j\} &= \frac{\alpha_j}{\hat{\alpha}}, \\ \text{var}\{\mu_j\} &= \frac{\alpha_j(\hat{\alpha} - \alpha_j)}{\hat{\alpha}^2(\hat{\alpha} + 1)}, \\ \text{cov}\{\mu_i, \mu_j\} &= -\frac{\alpha_i \alpha_j}{\hat{\alpha}^2(\hat{\alpha} + 1)}, \\ \mathbb{E}\{\log \mu_j\} &= \psi(\alpha_j) - \psi(\hat{\alpha}), \\ H(\boldsymbol{\mu}) &= -\sum_{j=1}^K (\alpha_j - 1) \{\psi(\alpha_j) - \psi(\hat{\alpha})\} + \log B(\boldsymbol{\alpha}),\end{aligned}$$

where

$$\begin{aligned}\hat{\alpha} &= \alpha_1 + \cdots + \alpha_K, \\ \psi(x) &= \frac{d}{dx} \log \Gamma(x), \quad (\text{digamma function}).\end{aligned}$$

Dirichlet Distribution (Cont'd)

- ▶ **Conjugate prior** for the **multinomial distribution** and a **generalization** of the **beta distribution**.
- ▶ Parameters α_j 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 the **finite density** everywhere, provided $\alpha_j \geq 1$ for all j .

13 / 25

14 / 25

Gamma Distribution

- ▶ A probability distribution over a **positive random variable** $\tau > 0$ governed by parameters $a > 0$ and $b > 0$:

$$\text{Gam}(\tau|a, b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau}.$$

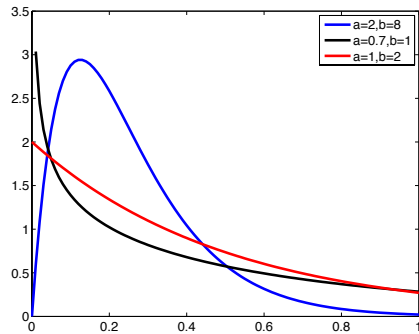
- ▶ Key statistics are given by

$$\begin{aligned}\mathbb{E}\{\tau\} &= \frac{a}{b}, \\ \text{var}\{\tau\} &= \frac{a}{b^2}, \\ \mathbb{E}\{\log \tau\} &= \psi(a) - \log b, \\ H(\tau) &= \log \Gamma(a) - (a - 1)\psi(a) - \log b + a.\end{aligned}$$

15 / 25

16 / 25

Gamma Distribution (Cont'd)

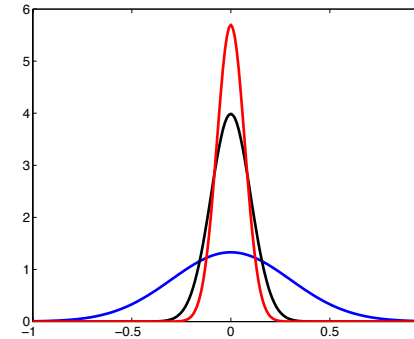


- ▶ Density is everywhere finite for $a \geq 1$.
- ▶ Reduces to a **exponential distribution** for $a = 1$.
- ▶ The **gamma distribution** is the **conjugate prior** for the **precision (inverse variance)** of a **univariate Gaussian**.

Univariate Gaussian Distribution

Distribution for continuous variables $x \in (-\infty, \infty)$, governed by two parameters μ and $\sigma^2 > 0$:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}.$$



- ▶ Conjugate prior for μ is the **Gaussian distribution**.
- ▶ Conjugate prior for $\tau = 1/\sigma^2$ is the **Gamma distribution**.
- ▶ Joint conjugate prior for both **unknown μ and τ** is the **Gaussian-gamma distribution**.

17 / 25

18 / 25

Multivariate Gaussian Distribution

- ▶ For a m -dimensional vector $\mathbf{x} \in \mathbb{R}^m$, the Gaussian is governed by a mean vector $\boldsymbol{\mu} \in \mathbb{R}^m$ and a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times m}$ that must be symmetric and positive-definite:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{m}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}.$$

- ▶ The inverse of the covariance matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ is referred to be as **precision matrix** which is also symmetric and positive definite.
- ▶ The entropy of \mathbf{x} is given by

$$H(\mathbf{x}) = \frac{1}{2} \log |\boldsymbol{\Sigma}| + \frac{m}{2} (1 + \log 2\pi).$$

19 / 25

Gaussian Identities

- ▶ Suppose that $\mathbf{x} = [\mathbf{x}_a^\top, \mathbf{x}_b^\top]^\top$ is **jointly Gaussian**,

$$\begin{aligned} \mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ab}^\top & \boldsymbol{\Sigma}_{bb} \end{bmatrix}\right) \\ &= \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ab}^\top & \boldsymbol{\Lambda}_{bb} \end{bmatrix}^{-1}\right). \end{aligned}$$

- ▶ **Marginal distributions** are $\mathbf{x}_a \sim \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$, $\mathbf{x}_b \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_{bb})$.
- ▶ **Conditional distributions** are:

$$\begin{aligned} p(\mathbf{x}_a | \mathbf{x}_b) &= \mathcal{N}\left(\boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab} \boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b), \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab} \boldsymbol{\Sigma}_{bb}^{-1} \boldsymbol{\Sigma}_{ab}^\top\right) \\ &= \mathcal{N}\left(\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b), \boldsymbol{\Lambda}_{aa}^{-1}\right), \\ p(\mathbf{x}_b | \mathbf{x}_a) &= \mathcal{N}\left(\boldsymbol{\mu}_b + \boldsymbol{\Sigma}_{ab}^\top \boldsymbol{\Sigma}_{aa}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_a), \boldsymbol{\Sigma}_{bb} - \boldsymbol{\Sigma}_{ab}^\top \boldsymbol{\Sigma}_{aa}^{-1} \boldsymbol{\Sigma}_{ab}\right) \\ &= \mathcal{N}\left(\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{bb}^{-1} \boldsymbol{\Lambda}_{ab}^\top(\mathbf{x}_a - \boldsymbol{\mu}_a), \boldsymbol{\Lambda}_{bb}^{-1}\right). \end{aligned}$$

20 / 25

Matrix Identities

- ▶ [Matrix inversion lemma](#)

$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}.$$

- ▶ [Inverse of partitioned matrix](#)

$$\mathbf{A} = \begin{bmatrix} \mathbf{P} & \mathbf{Q} \\ \mathbf{R} & \mathbf{S} \end{bmatrix}, \quad \mathbf{A}^{-1} = \begin{bmatrix} \tilde{\mathbf{P}} & \tilde{\mathbf{Q}} \\ \tilde{\mathbf{R}} & \tilde{\mathbf{S}} \end{bmatrix},$$

where sub-matrices are given by

$$\begin{aligned} \tilde{\mathbf{P}} &= \mathbf{P}^{-1} + \mathbf{P}^{-1}\mathbf{Q}\mathbf{M}\mathbf{R}\mathbf{P}^{-1}, \\ \tilde{\mathbf{Q}} &= -\mathbf{P}^{-1}\mathbf{Q}\mathbf{M}, \\ \tilde{\mathbf{R}} &= -\mathbf{M}\mathbf{R}\mathbf{P}^{-1}, \\ \tilde{\mathbf{S}} &= \mathbf{M}, \\ \mathbf{M} &= (\mathbf{S} - \mathbf{R}\mathbf{P}^{-1}\mathbf{Q})^{-1}. \end{aligned}$$

Wishart Distribution

- ▶ [Conjugate prior](#) for the [precision matrix](#) of a multivariate Gaussian:

$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu) = B(\mathbf{W}, \nu) |\boldsymbol{\Lambda}|^{(\nu-m-1)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda}) \right\},$$

where the normalization factor is given by

$$B(\mathbf{W}, \nu) = |\mathbf{W}|^{-\nu/2} \left(2^{\nu m/2} \pi^{m(m-1)/2} \prod_{i=1}^m \Gamma \left(\frac{\nu+1-i}{2} \right) \right)^{-1},$$

and the parameter ν is called the [number of degrees of freedom](#) of the distribution and is restricted to $\nu > m - 1$ to ensure the Gamma function in the normalization factor is well-defined.

- ▶ In one dimension, the Wishart reduces to the [Gamma distribution](#) $\text{Gam}(\lambda|a, b)$ with parameters $a = \nu/2$ and $b = 1/2W$.

More Gaussian Identities

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

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

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

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

where

$$\boldsymbol{\Phi} = (\boldsymbol{\Lambda} + \mathbf{A}^{\top}\mathbf{L}\mathbf{A})^{-1}.$$

21 / 25

22 / 25

Wishart Distribution (Cont'd)

Key statistics are given by

$$\begin{aligned} \mathbb{E}\{\boldsymbol{\Lambda}\} &= \nu\mathbf{W}, \\ \mathbb{E}\{\log|\boldsymbol{\Lambda}|\} &= \sum_{i=1}^m \psi \left(\frac{\nu+1-i}{2} \right) + m \log 2 + \log|\mathbf{W}|, \\ H(\boldsymbol{\Lambda}) &= -\log B(\mathbf{W}, \nu) - \frac{\nu-m-1}{2} \mathbb{E}\{\log|\boldsymbol{\Lambda}|\} + \frac{\nu m}{2}, \end{aligned}$$

where $\phi(\cdot)$ is the [digamma function](#).

23 / 25

24 / 25

Gaussian-Wishart Distribution

- ▶ **Conjugate prior** for a **multivariate Gaussian** $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ in which both the **mean** $\boldsymbol{\mu}$ and the **precision** $\boldsymbol{\Lambda}$ are **unknown**.
- ▶ Comprises the product of Gaussian distribution for $\boldsymbol{\mu}$, whose precision is proportional to $\boldsymbol{\Lambda}$ and a Wishart distribution over $\boldsymbol{\Lambda}$:

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

- ▶ For the particular case of a scalar x , it is equivalent to the **Gaussian-gamma distribution**:

$$p(\mu, \lambda|\mu_0, \beta, a, b) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1}) \text{Gam}(\lambda|a, b).$$