# STA 360: Reference Sheet for Distributions

# 1 Univariate discrete distributions

## 1.1 Uniform

**Notation**: $U\{a, b\}$

**Support**: $\mathcal{X} = \{a, a+1, \cdots, b\}$

**Probability mass function (p.m.f.)**:

$$\begin{aligned} p(x; a, b) &= \frac{1}{b-a+1} \quad \text{for} \quad x = a, a+1, \cdots, b \\ &\propto \mathbf{1}\{x \in \{a, a+1, \cdots, b\}\} \end{aligned}$$

**Parameters**:

- $a$: Lower bound ($a$ integer)
- $b$: Upper bound ($b > a$ integer)

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = \frac{a+b}{2}$$

**Variance**:

$$\mathrm{Var}(X) = \frac{(b-a+1)^2 - 1}{2}$$

**Notes**:

## 1.2 Bernoulli

**Notation**: $Bern(q)$

**Support**: $\mathcal{X} = \{0, 1\}$

**Probability mass function (p.m.f.)**:

$$p(x) = \begin{cases} 1 - q & \text{if } x = 0 \\ q & \text{if } x = 1 \end{cases}$$

**Parameters**:

- $q$: Probability of a success ($0 \leq q \leq 1$ real)

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = q$$

**Variance**:

$$\mathrm{Var}(X) = q(1 - q)$$

**Notes**:

- Models an experiment with two possible outcomes: a success or a failure.
- Building block for the binomial, geometric, and negative binomial distributions.

### 1.3 Binomial

**Notation**: $Bin(n, q)$

**Support**: $\mathcal{X} = \{0, 1, 2, \cdots, n\}$

**Probability mass function (p.m.f.)**:

$$p(x; n, q) = \binom{n}{x} q^x (1-q)^{n-x}, \quad x = 0, 1, \cdots, n$$

**Parameters**:

- $n$: Number of trials ($n$ positive integer)
- $q$: Probability of a success ($0 \leq q \leq 1$ real)

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = nq$$

**Variance**:

$$\mathrm{Var}(X) = nq(1-q)$$

**Notes**:

- Models the number of successes in an experiment with $n$ trials, where trials are i.i.d. $Bern(q)$ random variables.
- How do we interpret the coefficient $\binom{n}{x}$?
- $Bin(n, q)$ is the sum of $n$ i.i.d. $Bern(q)$.
- For large $n$, computation can be ill-conditioned and nasty (see Poisson distribution).

## 1.4 Geometric

**Notation**: $Geo(q)$

**Support**: (a) $\mathcal{X} = \{0, 1, 2, \cdots\}$, (b) $\mathcal{X} = \{1, 2, 3, \cdots\}$

**Probability mass function (p.m.f.)**:

$$\begin{array}{ll} \text{(a)} & p(x;q) = q(1-q)^x \propto (1-q)^x, \quad x = 0, 1, 2, \cdots \\ \text{(b)} & p(x;q) = q(1-q)^{x-1} \propto (1-q)^{x-1}, \quad x = 1, 2, 3, \cdots \end{array}$$

**Parameters**:

- $q$: Probability of a success ($0 \leq q \leq 1$ real)

**Visualization**:

**Mean**:

$$\text{(a)}: \quad \mathbb{E}(X) = \frac{1}{q} - 1, \quad \text{(b)}: \quad \mathbb{E}(X) = \frac{1}{q}$$

**Variance**:

$$\text{(a) and (b)}: \quad \text{Var}(X) = \frac{1-q}{q^2}$$

**Notes**:

- (a) models the number of *failures* needed to observe the first success, where trials are i.i.d. $Bern(q)$ random variables. (b) models the number of *trials* needed to observe the first success.

- Why are the means different for (a) and (b)? Why are the variances the same?

## 1.5 Poisson

**Notation**: $Poisson(\lambda)$

**Support**: $\mathcal{X} = \{0, 1, 2, \cdots\}$

**Probability mass function (p.m.f.)**:

$$p(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \propto \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \cdots$$

**Parameters**:

- $\lambda$: Rate parameter ($\lambda > 0$ real)

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = \lambda$$

**Variance**:

$$\mathrm{Var}(X) = \lambda$$

**Notes**:

- Widely-used model for count data; many extensions for more complex count data (e.g., Poisson point processes, Poisson regression, zero-inflated Poisson, etc.)
- *Justification 1*: Law of rare events
  - As $n \to \infty$, $q \to 0$, $nq \to \lambda$, the binomial distribution $Bin(n, q)$ converges to the Poisson distribution $Poisson(\lambda)$.
- *Justification 2*: Counts distribution under memoryless waiting times
  - Suppose waiting time between events follow i.i.d. $Exp(\lambda)$. Then the number of events in the time interval $[0, T]$ follow $Poisson(\lambda T)$.

## 1.6 Negative binomial

**Notation**: $NB(r, q)$

**Support**: (a) $\mathcal{X} = \{0, 1, 2, \cdots\}$, (b) $\mathcal{X} = \{r, r+1, r+2, \cdots\}$

**Probability mass function (p.m.f.)**:

$$\begin{aligned}
&\text{(a)} \quad p(x; r, q) = \binom{x+r-1}{x} q^r (1-q)^x \propto \binom{x+r-1}{x} (1-q)^x, \quad x = 0, 1, 2, \cdots \\
&\text{(b)} \quad p(x; r, q) = \binom{x-1}{r-1} q^r (1-q)^{x-r} \propto \binom{x-1}{r-1} (1-q)^{x-r}, \quad x = r, r+1, r+2, \cdots
\end{aligned}$$

**Parameters**:

- $q$: Probability of a success ($0 \leq q \leq 1$ real)
- $r$: Number of successes desired ($r$ positive integer)

**Visualization**:

**Mean**:

$$\text{(a):} \quad \mathbb{E}(X) = \frac{r}{q} - r, \quad \text{(b):} \quad \mathbb{E}(X) = \frac{r}{q}$$

**Variance**:

$$\text{(a) and (b):} \quad \mathrm{Var}(X) = \frac{r(1-q)}{q^2}$$

**Notes**:

- (a) models the number of *failures* needed to observe $r$ successes, where trials are independent Bernoulli random variables. (b) models the number of *trials* needed to observe $r$ successes.
- Why are the means different for (a) and (b)? Why are the variances the same?
- $NB(r, q)$ is the sum of $r$ i.i.d. $Geo(q)$.
- Good alternative to the Poisson distribution for count data when the variance of the data exceeds its average (*overdispersion*).

## 1.7 Hypergeometric

**Notation**: $HGeo(n, N, M)$

**Support**: $\mathcal{X} = \{(n - N + M)_+, \cdots, n \wedge M\}$ (note: $(z)_+ := \max(z, 0)$, $y \wedge z := \min(y, z)$)

**Probability mass function (p.m.f.)**:

$$p(x; n, N, M) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \propto \binom{M}{x}\binom{N-M}{n-x}, \quad x = (n - N + M)_+, \cdots, n \wedge M$$

**Parameters**:

- $N$: Number of elements in a finite population ($N$ positive integer)
- $M$: Number of "successes" in a finite population ($M < N$ positive integer)
- $n$: Number of samples without replacement ($n < N$ positive integer)

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = n\frac{M}{N}$$

**Variance**:

$$\text{Var}(X) = n\left(\frac{M}{N}\right)\left(1 - \frac{M}{N}\right)\left(\frac{N-n}{N-1}\right)$$

**Notes**:

- Models the number of "successes" when sampling $n$ elements from a finite population *without replacement*.
- How do we interpret the combination terms in the p.m.f.?
- How do the mean and variance of $HGeo(n, N, M)$ (sampling *without* replacement) compare with that for $Bin(n, q)$ with $q = M/N$ (sampling *with* replacement)?

# 2 Univariate continuous distributions

## 2.1 Uniform

**Notation**: $U[a, b]$

**Support**: $\mathcal{X} = [a, b]$

**Probability density function (p.d.f.)**:

$$\begin{aligned} p(x; a, b) &= \frac{1}{b-a}, \quad a \leq x \leq b \\ &\propto \mathbf{1}\{x \in [a, b]\} \end{aligned}$$

**Parameters**:

- $a$: Lower bound ($a$ real)
- $b$: Upper bound ($b > a$ real)

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = \frac{a+b}{2}$$

**Variance**:

$$\text{Var}(X) = \frac{(b-a)^2}{12}$$

**Notes**:

## 2.2 Normal

**Notation**: $N(\mu, \sigma^2)$

**Support**: $\mathcal{X} = \mathbb{R} := (-\infty, \infty)$

**Probability density function (p.d.f.)**:

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \propto e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad x \in \mathbb{R}$$

**Parameters**:

- $\mu$: Mean ($\mu$ real)
- $\sigma^2$: Variance ($\sigma^2 > 0$ real)

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = \mu$$

**Variance**:

$$\mathrm{Var}(X) = \sigma^2$$

**Notes**:

- Widely-used model for continuous data; many extensions for more complex data (e.g. Gaussian processes, mixture normal, multivariate normal, etc.)
- *Justification*: Central limit theorem
  - Suppose $X_1, \cdots, X_n$ are i.i.d. random variables with zero mean and variance $\sigma^2$. Then $\sqrt{n}\bar{X}_n \xrightarrow{d} N(0, \sigma^2)$.

## 2.3 Exponential

**Notation**: $Exp(\lambda)$

**Support**: $\mathcal{X} = (0, +\infty)$

**Probability density function (p.d.f.)**:

$$p(x; \lambda) = \lambda e^{-\lambda x} \propto e^{-\lambda x}, \quad x > 0$$

**Parameters**:

- $\lambda$: Rate parameter ($\lambda > 0$ real)

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = \frac{1}{\lambda}$$

**Variance**:

$$\text{Var}(X) = \frac{1}{\lambda^2}$$

**Notes**:

- Widely-used model for event times (e.g., time until next bus arrival, time until a radioactive particle decays, etc.)

- *Justification*: Memoryless property $P(X > t + s | X > t) = P(X > s)$.

  - Show this using conditional probabilities. Interpret this property when $X$ is the time until next bus arrival.

  $Exp(\lambda)$ is the *only* memoryless distribution over $\mathcal{X} = (0, +\infty)$.

- Suppose waiting time between events follow i.i.d. $Exp(\lambda)$. Then the number of events in a time interval $[0, T]$ follow $Poisson(\lambda T)$.

## 2.4 Beta

**Notation**: $Beta(a, b)$

**Support**: $\mathcal{X} = [0, 1]$

**Probability density function (p.d.f.)**:

$$p(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(\beta)} x^{a-1}(1-x)^{b-1} \propto x^{a-1}(1-x)^{b-1}, \quad 0 \leq x \leq 1$$

**Parameters**:

- $a$: Shape parameter ($a > 0$ real)
- $b$: Shape parameter ($b > 0$ real)

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = \frac{a}{a+b}$$

**Variance**:

$$\mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

**Notes**:

- Useful as a probabilistic model on proportions.
- If $a < b$, then $X \sim Beta(a, b)$ is more concentrated below 0.5; if $a > b$, then $X \sim Beta(a, b)$ is more concentrated above 0.5.
- If $X \sim Gamma(a, \theta)$ and $Y \sim Gamma(b, \theta)$, then $X/(X+Y) \sim Beta(a, b)$.
- If $X \sim U[0, 1]$ and $a > 0$, then $X^{1/a} \sim Beta(a, 1)$.

## 2.5 Chi-squared

**Notation**: $\chi^2(\nu)$

**Support**: $\mathcal{X} = (0, +\infty)$

**Probability density function (p.d.f.)**:

$$p(x; \nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} \propto x^{\nu/2-1} e^{-x/2}, \quad x > 0$$

**Parameters**:

- $\nu$: Degrees-of-freedom ($\nu$ positive integer)

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = \nu$$

**Variance**:

$$\mathrm{Var}(X) = 2\nu$$

**Notes**:

- If $X_1, \cdots, X_\nu$ are i.i.d. $N(0,1)$, then $\sum_{i=1}^{\nu} X_i^2 \sim \chi^2(\nu)$ (this is the basis behind F-tests in ANOVA, which are ratios of scaled, independent chi-squared distributions).

- The chi-squared distribution $X \sim \chi^2(\nu)$ is a special case of the Gamma distribution, namely, $Gamma(\nu/2, 1/2)$.

## 2.6 Gamma

**Notation**: $Gamma(a, b)$

**Support**: $\mathcal{X} = (0, +\infty)$

**Probability density function (p.d.f.)**:

$$p(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \propto x^{a-1} e^{-bx}, \quad x > 0$$

**Parameters**:

- $a$: Shape parameter ($a > 0$ real)
- $b$: Rate parameter ($b > 0$ real)

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = \frac{a}{b}$$

**Variance**:

$$\mathrm{Var}(X) = \frac{a}{b^2}$$

**Notes**:

- Flexible model for non-negative random variables (e.g., rainfall, age, etc.). Also widely used as a conjugate prior for precision (inverse variance) parameters.
- Includes some one-parameter distributions as special cases:
  - If $X \sim Exp(\lambda)$, then $X \sim Gamma(1, \lambda)$.
  - If $X \sim \chi^2(\nu)$, then $X \sim Gamma(\nu/2, 1/2)$.

## 2.7 Inverse-Gamma

**Notation**: $InvGamma(a, b)$

**Support**: $\mathcal{X} = (0, +\infty)$

**Probability density function (p.d.f.)**:

$$p(x; a, b) = \frac{b^a}{\Gamma(a)} x^{-a-1} e^{-b/x} \propto x^{-a-1} e^{-b/x}, \quad x > 0$$

**Parameters**:

- $a$: Shape parameter ($a > 0$ real)
- $b$: Scale parameter ($b > 0$ real)

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = \frac{b}{a-1} \quad \text{if} \quad a > 1$$

**Variance**:

$$\text{Var}(X) = \frac{b^2}{(a-1)^2(a-2)} \quad \text{if} \quad a > 2$$

**Notes**:

- Widely used as a conjugate prior for variance parameters.
- If $X \sim Gamma(a, b)$, then $1/X \sim InvGamma(a, b)$.

## 2.8 Laplacian

**Notation**: $Laplacian(\lambda)$

**Support**: $\mathcal{X} = \mathbb{R} := (-\infty, \infty)$

**Probability density function (p.d.f.)**:

$$p(x; \lambda) = \frac{\lambda}{2} e^{-\lambda|x|} \propto e^{-\lambda|x|}, \quad x \in \mathbb{R}$$

**Parameters**:

- $\lambda$: Rate parameter ($\lambda > 0$ real)

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = 0$$

**Variance**:

$$\mathrm{Var}(X) = \frac{2}{\lambda^2}$$

**Notes**:

- A two-sided extension of the exponential distribution.
- Used as a sparsity-inducing prior for Bayesian Lasso.

### 2.9 Pareto

**Notation**: $Pareto(m, \alpha)$

**Support**: $\mathcal{X} = [m, +\infty)$

**Probability density function (p.d.f.)**:

$$p(x; m, \alpha) = \frac{\alpha m^\alpha}{x^{\alpha+1}} \propto \frac{1}{x^{\alpha+1}}, \quad x \geq m$$

**Parameters**:

- $m$: Scale parameter ($m > 0$ real)
- $\alpha$: Shape parameter ($\alpha > 0$ real)

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = \frac{\alpha m}{\alpha - 1} \quad \text{if} \quad \alpha > 1$$

**Variance**:

$$\text{Var}(X) = \frac{m^2 \alpha}{(\alpha - 1)^2 (\alpha - 2)} \quad \text{if} \quad \alpha > 2$$

**Notes**:

- Widely used as a model for wealth distribution among individuals. The Pareto distribution implicitly encodes the *Pareto principle*: a larger portion of wealth is owned by a smaller percentage of people in a society.
- Also useful for modeling data where an equilibrium is found in the distribution of the "small" to the "large" (e.g., insurance losses, size of human settlements, etc.)

## 2.10 Lognormal

**Notation**: $Lognormal(\mu, \sigma^2)$

**Support**: $\mathcal{X} = (0, +\infty)$

**Probability density function (p.d.f.)**:

$$p(x; \mu, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} \propto \frac{1}{x} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \quad x > 0$$

**Parameters**:

- $\mu$ real
- $\sigma^2 > 0$ real

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = e^{\mu + \frac{\sigma^2}{2}}$$

**Variance**:

$$\mathrm{Var}(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$$

**Notes**:

- If $X \sim Lognormal(\mu, \sigma^2)$, then $\ln X \sim N(\mu, \sigma^2)$.
- Models natural growth processes / phenomena. The idea is that many phenomena are driven by the multiplicative accumulation of small changes, which become additive on a log-scale. If such changes are i.i.d., the central limit theorem says their sum is approximately normal, so the original phenomena is approximately lognormal (after a back-transformation).
- Widely used in financial option pricing, neuron firing rates, size of living tissues, etc.

## 2.11 Weibull

**Notation**: $Weibull(\lambda, k)$

**Support**: $\mathcal{X} = (0, +\infty)$

**Probability density function (p.d.f.)**:

$$p(x; \lambda, k) = k\lambda \left(x\lambda\right)^{k-1} e^{-(x\lambda)^k} \propto x^{k-1} e^{-(x\lambda)^k}, \quad x > 0$$

**Parameters**:

- $\lambda$: Rate parameter ($\lambda > 0$ real)
- $k$: Shape parameter ($k > 0$ real)

**Visualization**:

**Mean**:

$$\mathbb{E}(X) = \frac{1}{\lambda}\Gamma\left(1 + \frac{1}{k}\right)$$

**Variance**:

$$\mathrm{Var}(X) = \frac{1}{\lambda^2}\left[\Gamma\left(1 + \frac{2}{k}\right) - \left(\Gamma\left(1 + \frac{1}{k}\right)\right)^2\right]$$

**Notes**:

- If $W \sim Exp(\lambda)$, then $W^k \sim Weibull(\lambda, k)$.
- Widely used in survival analysis and reliability engineering, to model the "time-to-failure" of a component:
  - $k < 1$: failure rate decreases over time (failures more likely initially)
  - $k = 1$: failure rate constant in time (memoryless)
  - $k > 1$: failure rate increases in time (failures more likely as time goes on; an "aging" process)

# 3 Multivariate distributions

## 3.1 Categorical

**Notation**: $Categorical(\mathbf{p})$, $\mathbf{p} := (p_1, \cdots, p_k)$

**Support**: $\mathcal{X} = \{\mathbf{x}_k \in \{0,1\} \ : \ \sum_{k=1}^{K} x_k = 1\}$

**Probability mass function (p.m.f.)**:

$$p(\mathbf{x}; \mathbf{p}) = p_1^{x_1} \cdots p_K^{x_K}, \quad x_k \in \{0,1\}, \quad \sum_{k=1}^{K} x_k = 1$$

**Parameters**:

- $\mathbf{p}$: Vector of probabilities corresponding to the $K$ categories.

**Visualization**:

**Mean**:

$$\mathbb{E}(X_k) = p_k, \quad k = 1, \cdots, K$$

**Variance**:

$$\begin{aligned} \mathrm{Var}(X_k) &= p_k(1 - p_k), \quad k = 1, \cdots, K \\ \mathrm{Cov}(X_k, X_l) &= -p_k p_l, \quad i, j = 1, \cdots, K, \quad i \neq j \end{aligned}$$

**Notes**:

- Models the sampling (with replacement) of one category from $K$ possible categories with probabilities $\mathbf{p}$.
- $\mathbf{x} = (x_1, \cdots, x_K)$ represents the number of times a category has been selected. Note that only one entry is a '1' (the category selected); all other entries are '0's.
- Multivariate extension of the Bernoulli distribution:
  - If $\mathbf{X} \sim Categorical(\mathbf{p})$, then $X_k \sim Bernoulli(p_k)$, $k = 1, \cdots, K$.
  - Are $X_1$ and $X_2$ correlated?

### 3.2 Multinomial

**Notation**: $Multinomial(n, \mathbf{p})$

**Support**: $\mathcal{X} = \{x_k \in \{0, \cdots, n\} \ : \ \sum_{k=1}^K x_k = n\}$

**Probability mass function (p.m.f.)**:

$$p(\mathbf{x}; n, \mathbf{p}) = \frac{n!}{x_1! \cdots x_K!} p_1^{x_1} \cdots p_K^{x_K}, \quad x_k \in \{0, \cdots, n\}, \quad \sum_{k=1}^K x_k = n$$

**Parameters**:

- $n$: Number of trials ($n$ positive integer)
- $\mathbf{p}$: Vector of probabilities corresponding to the $K$ categories.

**Visualization**:

**Mean**:

$$\mathbb{E}(X_k) = np_k, \quad k = 1, \cdots, K$$

**Variance**:

$$\begin{aligned} \text{Var}(X_k) &= np_k(1 - p_k), \quad k = 1, \cdots, K \\ \text{Cov}(X_k, X_l) &= -np_k p_l, \quad i, j = 1, \cdots, K, \quad i \neq j \end{aligned}$$

**Notes**:

- Models the sampling (with replacement) of $n$ categories from $K$ possible categories with probabilities $\mathbf{p}$.
- $\mathbf{x} = (x_1, \cdots, x_K)$ represents the number of times a category has been selected. The vector should sum to $n$ (since $n$ categories are sampled).
- Multivariate extension of the Binomial distribution:
  - If $\mathbf{X} \sim Multinomial(n, \mathbf{p})$, then $X_k \sim Binomial(n, p_k)$, $k = 1, \cdots, K$.
  - Are $X_1$ and $X_2$ correlated?
- $\mathbf{X} \sim Multinomial(1, \mathbf{p}) \Rightarrow \mathbf{X} \sim Categorical(\mathbf{p})$

### 3.3 Multivariate normal

**Notation**: $MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

**Support**: $\mathcal{X} = \mathbb{R}^d$
**Probability density function (p.d.f.)**:

$$p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2}\det(\boldsymbol{\Sigma})^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \propto e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \quad \mathbf{x} \in \mathbb{R}^d$$

**Parameters**:

- $\boldsymbol{\mu}$: Mean vector ($\mu_i$ real)
- $\boldsymbol{\Sigma}$: Covariance matrix ($\boldsymbol{\Sigma}$ symmetric, positive-definite)

**Visualization**:

**Mean**:

$$\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu} \quad (\text{equivalently, } \mathbb{E}(X_i) = \mu_i,\ i = 1, \cdots, d)$$

**Variance**:

$$\text{Var}(\mathbf{X}) = \boldsymbol{\Sigma} \quad (\text{equivalently, } \text{Cov}(X_i, X_j) = \Sigma_{i,j},\ i, j = 1, \cdots, d)$$

**Notes**:

- Widely-used model for multivariate continuous data
- *Justification*: (Multivariate) Central limit theorem
- Multivariate extension of the normal distribution:
  - If $\mathbf{X} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $X_i \sim N(\mu_i, \Sigma_{i,i})$, $i = 1, \cdots, d$ (note: $\Sigma_{i,i}$ is the $i$-th entry on the diagonal of $\boldsymbol{\Sigma}$).
  - Are $X_1$ and $X_2$ correlated? Are they independent?

## 3.4 Dirichlet

**Notation**: $\mathit{Dirichlet}(\boldsymbol{\alpha})$

**Support**: $\mathcal{X} = \{x_k \in [0,1] \ : \ \sum_{k=1}^{K} x_k = 1\}$

**Probability density function (p.d.f.)**:

$$p(\mathbf{x};\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}\prod_{k=1}^{K} x_k^{\alpha_k-1} \propto \prod_{k=1}^{K} x_k^{\alpha_k-1}, \quad x_k \in [0,1], \quad \sum_{k=1}^{K} x_k = 1$$

**Parameters**:

- $\boldsymbol{\alpha}$: Vector of concentration parameters ($\alpha_i > 0$ real)

**Visualization**:

**Mean**:

$$\mathbb{E}(X_i) = \frac{\alpha_i}{\sum_{k=1}^{K}\alpha_k}$$

**Variance**:

$$\mathrm{Var}(X_i) = \frac{\gamma_i(1-\gamma_i)}{\sum_{k=1}^{K}\alpha_k + 1}, \quad \gamma_i := \frac{\alpha_i}{\sum_{k=1}^{K}\alpha_k}$$

**Notes**:

- Useful as a probabilistic model on a vector of proportions (summing to 1).
- Multivariate extension of the beta distribution:
  - If $\mathbf{X} \sim \mathit{Dirichlet}(\boldsymbol{\alpha})$, then $X_i \sim \mathit{Beta}(\alpha_i, \sum_{k=1}^{K}\alpha_k - \alpha_i)$, $i = 1, \cdots, K$.

# 4 Matrix-variate distributions

## 4.1 Wishart

**Notation**: $W(\Psi, \nu)$

**Support**: $\mathcal{X} = \{\mathbf{\Sigma} \in \mathbb{R}^{d\times d} \ : \ \mathbf{\Sigma} \text{ symmetric, positive-definite}\}$

**Probability density function (p.d.f.)**:

$$
\begin{aligned}
p(\mathbf{\Sigma}; \Psi, \nu) &= \frac{1}{2^{\nu d/2}\det(\Psi)^{\nu/2}\Gamma_d(\nu/2)}\det(\mathbf{\Sigma})^{(\nu-d-1)/2}e^{-\mathrm{tr}(\Psi^{-1}\mathbf{\Sigma})/2} \\
&\propto \det(\mathbf{\Sigma})^{(\nu-d-1)/2}e^{-\mathrm{tr}(\Psi^{-1}\mathbf{\Sigma})/2}, \quad \mathbf{\Sigma} \text{ sym. p.d.}
\end{aligned}
$$

**Parameters**:

- $\Psi$: scale matrix ($\Psi \in \mathbb{R}^{d\times d}$ p.d.)
- $\nu$: degrees-of-freedom ($\nu > d-1$ real)

**Visualization**:

**Mean**:

$$\mathbb{E}(\mathbf{\Sigma}_{i,j}) = \nu\Psi_{i,j}, \quad i,j = 1, \cdots, d$$

**Variance**:

$$\mathrm{Var}(\mathbf{\Sigma}_{i,j}) = \nu(\Psi_{i,j}^2 + \Psi_{i,i}\Psi_{j,j}), \quad i,j = 1, \cdots, d$$

**Notes**:

- Useful as a probabilistic model on inverse covariance matrices (which must be symmetric and positive-definite).
- Matrix-variate extension of the Gamma distribution.

## 4.2 Inverse-Wishart

**Notation**: $IW(\Psi, \nu)$

**Support**: $\mathcal{X} = \{\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d} \ : \ \boldsymbol{\Sigma} \text{ symmetric, positive-definite}\}$

**Probability density function (p.d.f.)**:

$$\begin{aligned} p(\boldsymbol{\Sigma}; \Psi, \nu) &= \frac{\det(\Psi)^{\nu/2}}{2^{\nu d/2}\Gamma_d(\nu/2)} \det(\boldsymbol{\Sigma})^{-(\nu+d+1)/2} e^{-\mathrm{tr}(\Psi \boldsymbol{\Sigma}^{-1})/2} \\ &\propto \det(\boldsymbol{\Sigma})^{-(\nu+d+1)/2} e^{-\mathrm{tr}(\Psi \boldsymbol{\Sigma}^{-1})/2}, \quad \boldsymbol{\Sigma} \text{ sym. p.d.} \end{aligned}$$

**Parameters**:

- $\Psi$: scale matrix ($\Psi \in \mathbb{R}^{d \times d}$ p.d.)
- $\nu$: degrees-of-freedom ($\nu > d - 1$ real)

**Visualization**:

**Mean**:

$$\mathbb{E}(\boldsymbol{\Sigma}_{i,j}) = \frac{1}{\nu - d - 1} \Psi, \quad i, j = 1, \cdots, d \quad \text{for } \nu > d + 1$$

**Notes**:

- Useful as a probabilistic model on covariance matrices (which must be symmetric and positive-definite).
- Matrix-variate extension of the Inverse-Gamma distribution.
- If $\boldsymbol{\Sigma} \sim W(\Psi, \nu)$, then $\boldsymbol{\Sigma}^{-1} \sim IW(\Psi^{-1}, \nu)$.