Probability Notes

Mark DiBattista

April 1, 2019

Abstract

Probability.

1 Suggested Resource Materials

Useful source texts:

- Probability/Statistics, intermediate (probability sections are better than statistics): *Statistical Inference*, Casella & Berger
- Probability, advanced: *Probability and Measure*, Billingsley

Throughout the text the acronym, *LAN*, refers to the companion writeup, *Linear Algebra Notes*, from which information is referenced by chapter and/or numbered equation.

2 Probability Preliminaries

Practical concerns in probability – those related to the calculation of probabilities and likelihoods that arise through models of physical phenomena – are addressed completely through operations on density or cumulative functions defined over real-valued spaces. For the purposes of compact presentation of mathematical relations, however, it is beneficial to ground the practical expressions in terms of an abstract theory, which covers the notions of the following:

- An **abstract probability space**, within which a probability measure is defined over collections of events,
- A **state space**, over which a probability density function captures the frequency at which events are mapped to real numbers,
- A **random variable** , an invertible function that allows movement from one space to the other as necessary.

It is usually the case that a given statement or relation is expressed most clearly in the abstract probability space, while specific realizations are carried out as operations on probability distributions in the state space.

2.1 Probability Spaces, State Spaces and Random Variables

Probability theory models the results of physical processes as individual elements, known as **outcomes**, and quantifies carefully selected aggregations of outcomes, known as **events**, by assigning real values to each. Both the collection of aggregations and assigned values are controlled, and must obey the following constraints:

- All events are assigned values between zero and unity, inclusive;
- The event of no outcomes is assigned a value of zero; the event of all outcomes is assigned a value of unity;
- For each event the *complement* aggregation – all outcomes are assigned to one aggregation or the other – is also an event;
- For two *disjoint* events – those that share no common outcome – the aggregation of both collections of outcomes is an event, and the value assigned to the joint aggregation is the sum of the values assigned to each separately;
- For a *countaby infinite* number of disjoint events, the *limit* of the aggregation of all collections is an event, and the value assigned to the joint aggregation is the *limit* of the sum of the values assigned to each separately.

2.1.1 Set Collections

Within probability theory individual outcomes are taken as featureless points distinguished by name only, events are interpreted as formal sets of outcomes, and values are assigned by a set function that maps events to the unit interval. The constraints on event formation, described in the itemized list above, is concisely expressed as the closure of a collection of sets, known as a ***σ*-algebra**, under operations of complementation and countable union:

$$\mathcal{S} \text{ is a } \sigma\text{-algebra} \Rightarrow \begin{cases} \emptyset \in \mathcal{S} \\ A \in \mathcal{S} \Rightarrow A^{\mathsf{c}} \in \mathcal{S} \\ A_{n\in\mathbb{N}} \in \mathcal{S} \Rightarrow \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{S} \end{cases} . \tag{1}$$

Notice that these rules imply that the σ-algebra is closed under countable intersection as well:

$$A_{n\in\mathbb{N}} \in \mathcal{S} \Rightarrow A^{\mathsf{c}}_{n\in\mathbb{N}} \Rightarrow \bigcup_{n\in\mathbb{N}} A^{\mathsf{c}}_n \in \mathcal{S} \Rightarrow \bigcap_{n\in\mathbb{N}} A_n \in \mathcal{S}, \tag{2}$$

since the complementation of countable union is countable intersection.

By the nature of inclusion and closure defined in (1), if events are organized into two distinct σ-algebras, then one must include the other. For a given collection of sets, $\mathcal{A}$, many σ-algebras may contain all its members, and the *smallest* σ-algebra that contains $\mathcal{A}$, formed from the intersection of all such σ-algebras, is indicated by $\sigma(\mathcal{A})$. In other words, if the full list of events is *partially* defined by inclusion in $\mathcal{A}$, the designation, $\sigma(\mathcal{A})$, specifies the smallest collection of *all consistent* events.

2.1.2 Set Functions

Within probability theory the values assigned to events must never decrease with respect to increasing combination of events. The constraints on set measures, described in the itemized list above, is captured by the requirements of a **subadditive measure**:

$$\mathbb{P} \text{ is a subadditive measure} \Rightarrow \begin{cases} \mathbb{P}\emptyset = 0 \\ A_1 \cap A_2 = \emptyset \Rightarrow \mathbb{P}(A_1 \cup A_2) = \mathbb{P}A_1 + \mathbb{P}A_2 \\ A_i \cap A_j = \emptyset, i \neq j \in \mathbb{N} \Rightarrow \mathbb{P}(\bigcup_{n\in\mathbb{N}}) = \sum_{n\in\mathbb{N}} \mathbb{P}A_n \end{cases} . \tag{3}$$

The specification of *sub*additivity comes from the application of the measure function to arbitrary sets:

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}((A - B) \cup (B - A) \cup (A \cap B)) \\ &\leq \mathbb{P}((A - B) \cup (A \cap B)) + \mathbb{P}((B - A) \cup (A \cap B)) = \mathbb{P}A + \mathbb{P}B, \end{aligned} \tag{4}$$

which provides an alternative basis for the definition.

2.1.3 Measurability of Sets and Functions

The definitions of σ-algebra in §2.1.1 and subadditive measure in §2.1.2 are joined in the notion of **measurability**. Indeed, for a given universe of all outcomes, Ω, the set of all events, $\mathcal{F}$, and a subadditive set function, $\mathbb{P}$, we have the definitions,

$$\left.\begin{array}{lr} \text{A universe of elements,} & \Omega \\ \text{A } \sigma\text{-algebra of sets,} & \mathcal{F} \end{array}\right\} \quad A \in \mathcal{F} \Rightarrow A \subset \Omega \text{ then } \Omega \text{ is } \mathcal{F}\text{-measurable;} \tag{5}$$

and

$$\left.\begin{array}{lr} \text{A } \sigma\text{-algebra of sets,} & \mathcal{F} \\ \text{A subadditive measure,} & \mathbb{P} \end{array}\right\} \quad \mathbb{P} : \mathcal{F} \to [0, 1] \text{ then } \mathcal{F} \text{ is } \mathbb{P}\text{-measurable.} \tag{6}$$

Thus, measurability is a concept that applies in a closely coupled fashion to both collections of sets and to subadditive set functions.

2.1.4 Probability and State Spaces

A **Probability Space** is defined as a triple,

$$\text{probability space } \to (\Omega, \mathcal{F}, \mathbb{P}), \tag{7}$$

consisting of the set of all outcomes, the set of all events and a set function that assigns each event a real value in the unit interval. In particular, the set of all events is a σ-algebra, and the set function is a **probability measure**, which in addition to the requirements of subadditivity listed in (3), is bounded by unity:

$$A \subset \Omega \Rightarrow 0 = \mathbb{P}\emptyset \leq \mathbb{P}A \leq \mathbb{P}\Omega = 1. \tag{8}$$

The key point is that the set of all outcomes, the set of all events, and the probability measure are linked by measurability, and the arrangment is sufficiently flexible to model the effects of chained 'and' and 'or' conditions on probabilistic statements (as well as the convergence properties of countably infinite may such statements), while also sufficiently restrictive to prevent the application to sets to which consistent probability values cannot be assigned, and cannot arise as natural problems of physical origin. The properties of the probability triple are summarized in the table:

Set of all outcomes,	Ω	Individual outcome,	ω	$\Omega = \{\omega_{\lambda\in\Lambda}\}$	(9)
σ-algebra of all events,	$\mathcal{F}$	Individual event,	$A \subset \Omega$	$\mathcal{F} = \{A_{\lambda\in\Lambda}\}$	(10)
		Ω is-$\mathcal{F}$-measurable			
Probability measure,	$\mathbb{P}$	$\mathcal{F}$ is $\mathbb{P}$-measurable		$\mathbb{P} : \mathcal{F} \to [0, 1]$	(11)

The index and index set, λ and Λ, respectively, in the table reference (possibly) uncountably infinite members in both the sets of outcomes and events.

Although a probability space provides a rigorous structure for modeling uncertainty in physical problems, it is unwieldy for performing calculations to address specific questions on probabilities of outcomes, many of which are indifferent to the detailed, computationally indistinguishable combinations of outcomes. The associated **State Space** is a kind of probability triple introduced to satisfy this,

$$\text{state space } \to (\mathbb{R}, \mathcal{B}, \mu). \tag{12}$$

Here, the universe of outcomes is taken as the real number line, $\mathbb{R}$, whose elements are collected into **Borel sets**, which are contained within the σ-algebra of events generated by the set of real intervals with rational endpoints – a countable set denoted as $\mathcal{J}$:

$$\mathcal{B} = \sigma(\mathcal{J}). \tag{13}$$

Finally, the set functions, $\mathbb{P}$ and μ, are both probability measures.

2.1.5 Random Variables

The probability space and the state space, which serve as kinds of abstract and realized domains of chance phenomena, are linked through a mapping, X, known as a **random variable**,

$$X : \Omega \to \mathbb{R} \quad \begin{cases} \text{Realization of the random variable:} & \omega \in \Omega \Rightarrow X(\omega) = x \\ \text{Borel sets pull back to measurable collections of events:} & X^{-1} : \mathcal{B} \to \mathcal{F} \\ \text{The probability `law of X':} \qquad \left.\begin{matrix} \mu_X : \mathcal{B} \to [0,1] \\ B \in \mathcal{B} \end{matrix}\right\} & \Rightarrow \mu_X B = \mathbb{P}_X X^{-1} B \end{cases} \tag{14}$$

The key point is the random variable, X, maps events in the probability space into events in the state space, and the values of the respective probability measures are identical for all events. The probability space is the more natural space to pose questions, the state space is the more natural to derive numerical results, and the random variable ensures that the two domains are consistently aligned in all cases.

Also note the slight change in symbol for the probability e measure, which has taken a subscript, $\mathbb{P}_X$. This is intended to indicate that the probability measure is restricted to sets in the probability space that are pulled back from Borel sets in the state space, as defined through the mapping, X.

2.1.6 The Probability Density and Cumulative Distribution Functions

The probability measure induced by the random variable, X, is expressed as a **cumulative distribution function**, F_X, which is the value assigned to the infinite half-open interval:

$$\left.\begin{array}{l} B = [-\infty, x) \\ A = X^{-1}B = \{\omega : X(\omega) < x\} \end{array}\right\} \Rightarrow \mathbb{P}_X A = \mu_X B \equiv F_X(x) \tag{15}$$

Applying the fundamental theorem of calculus, we can derive an equivalent expression in terms of a related **probability density function**, $p(x)$. The two functions are related by the standard operations,

$$\textbf{Probability density function:} \qquad p(x) = \frac{d}{dx}F(x) \tag{16}$$

$$\textbf{Cumulative distribution function:} \qquad F(x) = \int_{-\infty}^{x} p(x)\,dx \tag{17}$$

As a practical matter, 'probability distributions' are specified by attaching the random variable, X, to a formula for the probability density or cumulative distribution function.

2.1.7 The Expectation Operator

There is really only one operation in probability theory: integration of the entire probability space, organized over sets defined by the mapping, X, and weighted by functions of the random variable or by functions of the probability measure itself. This operation is known as **expectation**, and is designated by the operator, E. As an example, the mean is given by the expectation of X:

$$\mathbb{E}X = \int_\Omega X d\mathbb{P}_X = \int_{-\infty}^{\infty} x\,dF_X = \int_{-\infty}^{\infty} x\,p_X(x)\,dx. \tag{18}$$

Note also that probability calculations can be expressed in terms of the expectation operator as well,

$$\mathbb{P}A = \mathbb{E}\,1_A, \tag{19}$$

for which the operator, 1_A, is the indicator function for the set, A.

2.2 Standard Nomenclature

$$\left.\begin{array}{ll} A: & \text{Event} \\ X, Y: & \text{Random Variables} \end{array}\right\} \Rightarrow \begin{array}{ll} \text{Probability:} & \mathbb{E}\,1_A \\ \text{Mean:} & \mathbb{E}X \\ \text{Variance:} & \mathbb{V}X = \mathbb{E}(X - \mathbb{E}X)^2 \\ \text{Covariance:} & \mathbb{C}(X,Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) \\ \text{Information:} & \mathbb{I}X = -\ln \mathbb{P}_X \\ \text{Entropy:} & \mathbb{H}X = \mathbb{E}\mathbb{I}X \end{array} \tag{20}$$

3 Moments

3.1 Discrete

3.2 Continuous

3.2.1 Univariate

$$\mathbb{E}X \equiv \mu_X = \int_{-\infty}^{\infty} x \, p_X(x) \, dx \tag{21}$$

$$\mathbb{E}(X - \mathbf{E}X)^n \tag{22}$$

Variance

$$\begin{aligned} \mathbb{V}X \equiv \sigma_X^2 = \mathbb{E}(X - \mathbb{E}X)^2 &= \int_{-\infty}^{\infty} (x - \mu_X)^2 \, p(x) \, dx && (23) \\ = \mathbb{E}X^2 - (\mathbb{E}X)^2 &= \int_{-\infty}^{\infty} x^2 \, p(x) \, dx - \mu_X^2 && (24) \end{aligned}$$

Covariance

$$\begin{aligned} \mathbb{C}(X, Y \equiv \sigma_{XY} = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) \, p(x, y) \, dx \, dy && (25) \\ = \mathbb{E}XY - \mathbb{E}X \, \mathbb{E}Y &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \, p(x, y) \, dx \, dy - \mu_X \mu_Y && (26) \end{aligned}$$

3.2.2 Multivariate Moments

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \Rightarrow \left\{ \begin{aligned} \mathbb{E}\mathbf{X} \quad &= \begin{pmatrix} \mathbb{E}X_1 \\ \vdots \\ \mathbb{E}X_n \end{pmatrix} \\ \mathbb{V}\mathbf{X} \quad &= \mathbb{E}(\mathbf{X} - \mathbb{E}\mathbf{X}) \, \mathbb{E}(\mathbf{X} - \mathbb{E}\mathbf{X})^\top = \mathbb{E}\mathbf{X}\mathbf{X}^\top - \mathbb{E}\mathbf{X} \, \mathbb{E}\mathbf{X}^\top \\ &= \begin{pmatrix} \mathbb{V}X_1 & \cdots & \mathbb{C}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \mathbb{C}(X_n, X_1) & \cdots & \mathbb{V}X_n \end{pmatrix} \end{aligned} \right. \tag{27}$$

3.3 Sampled

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix} = \begin{pmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ \downarrow & & \downarrow \end{pmatrix} \tag{28}$$

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 & \cdots & \bar{x}_n \end{pmatrix} \to \bar{x}_j = \frac{1}{m} \sum_{i=1}^{m} x_{ij} \tag{29}$$

$$X^\top \mathbf{1}_m = m \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_n \end{pmatrix} \Rightarrow X^\top \mathbf{1}_m (\mathbf{1}_m^\top \mathbf{1}_m)^{-1} \mathbf{1}_m^\top X = X^\top P_{\mathbf{1}_m} X = m \begin{pmatrix} \bar{x}_1^2 & \cdots & \bar{x}_1 \bar{x}_n \\ \vdots & \ddots & \vdots \\ \bar{x}_n \bar{x}_1 & \cdots & \bar{x}_n^2 \end{pmatrix} \tag{30}$$

$$m\Sigma = X^\top X - X^\top P_{\mathbf{1}_m} X = X^\top I_m X - X^\top P_{\mathbf{1}_m} X = X^\top (I_m - P_{\mathbf{1}_m}) X \tag{31}$$

4 Norms and Inequalities

4.1 Cauchy-Schwartz Inequality

$$|\langle \mathbf{x}_1, \mathbf{x}_2 \rangle|^2 \leq \langle \mathbf{x}_1, \mathbf{x}_1 \rangle \cdot \langle \mathbf{x}_2, \mathbf{x}_2 \rangle \tag{32}$$

$$\mathbb{C}(X_1, X_2) \leq \mathbb{V}X_1 \cdot \mathbb{V}X_2 \tag{33}$$

4.2 Chebyshev's Inequality

$$\begin{aligned} \mathbb{P}\{g(X) \geq r\} \leq \frac{\mathbb{E}g(X)}{r} = \int_D g(x) f(x)\, dx \geq \int_{\{g(x) \geq r\}} g(x) f(x)\, dx \\ \geq r \int_{\{g(x) \geq r\}} f(x)\, dx = r\mathbb{P}\{g(X) \geq r\}. \end{aligned} \tag{34}$$

4.3 Jensen's Inequality

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X) \tag{35}$$

convex function

$$\phi(tx_1 + (1-t)x_2) \leq t\phi(x_1) + (1-t)\phi(x_2) \tag{36}$$

$$ax + b \leq \phi(x) \tag{37}$$

$$\begin{aligned} \mathbb{E}\phi(X) = \int_I \phi(x) p(x)\, dx \geq \int_I (ax+b) p(x)\, dx = \\ a \int_I x p(x)\, dx + b \int_I p(x)\, dx = ax_0 + b = \phi(x_0) = \phi(\mathbb{E}X). \end{aligned} \tag{38}$$

5 Operators

5.1 Exponentiated Operators

5.1.1 Moment-Generating Functions

$$M_X(t) \equiv \mathbb{E}e^{tX} \tag{39}$$

$$M_X(t) \equiv \sum_n = 1^\infty \frac{\mathbb{E}X^n}{n!} t^n \tag{40}$$

$$\mathbb{E}X^n = \frac{d^n}{dt^n} M_X(t)|_{t=0} = M_X^{(n)}(0) \tag{41}$$

X and Y independent

$$M_{X+Y}(t) = \mathbb{E}e^{t(X+Y)} = \mathbb{E}e^{tX}e^{tY} = \mathbb{E}e^{tX}\mathbb{E}e^{tY} = M_X(t)M_Y(t) \tag{42}$$

$$M_{cX}(t) = \mathbb{E}e^{ctX} = M_X(ct) \tag{43}$$

5.1.2 Cumulants

$$K_X(t) = \ln M_X(t) \tag{44}$$

Since $\ln(1+x) = t - \frac{t^2}{2} + \cdots$

$$K_X(t) = \left(t\mathbb{E}X + \frac{t^2}{2}\mathbb{E}X^2 + \cdots \right) + \frac{1}{2}\left(t\mathbb{E}X + \cdots\right)^2 + \cdots \tag{45}$$

$$= t\mathbb{E}X + \frac{t^2}{2}\left((\mathbb{E}X)^2 - \mathbb{E}X^2 \right) + \cdots \tag{46}$$

For X and Y independent

$$K_{X+Y}(t) = K_X(t) + K_Y(t) \tag{47}$$

$$K_{cX}(t) = K_X(ct) \tag{48}$$

5.1.3 Characteristic Functions

Fourier transform of probability density function

$$\phi_X(t) \equiv \mathbb{E}e^{itX} \tag{49}$$

for X and Y independent

$$\phi_{X+Y}(t) = \mathbb{E}e^{it(X+Y)} = \mathbb{E}e^{itX}e^{itY} = \mathbb{E}e^{itX}\mathbb{E}e^{itY} = \phi_X(t)\phi_Y(t) \tag{50}$$

$$\phi_{cX}(t) = \mathbb{E}e^{ictX} = \phi_X(ct) \tag{51}$$

let $Z = X + Y$ be the sum of two independent random variables

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) \Rightarrow p_{X+Y}(z) = \int_{-\infty}^{\infty} p_X(x)p_Y(z-x)\,dx \tag{52}$$

5.1.4 Extensions to Random Vectors

$$\left.\begin{array}{l} X \to \mathbf{X} \\ t \to \mathbf{t} \end{array}\right\} \Rightarrow \begin{cases} M_{\mathbf{X}}(\mathbf{t}) \equiv \mathbb{E}e^{\mathbf{t}^\top \mathbf{X}} \\ K_{\mathbf{X}}(\mathbf{t}) \equiv \ln M_{\mathbf{X}}(\mathbf{t}) \\ \phi_{\mathbf{X}}(\mathbf{t}) \equiv \mathbb{E}e^{i\mathbf{t}^\top \mathbf{X}} \end{cases} \tag{53}$$

5.2 Transformations

5.2.1 General Transformation

$$Y = g(X) \tag{54}$$

$$\text{increasing function, } g : F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq Y = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) \tag{55}$$

$$\text{decreasing function, } g : F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq Y = \mathbb{P}(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)) \tag{56}$$

$$f_Y(y) = \frac{d}{dy} F_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \tag{57}$$

5.2.2 Scale-location Adjustment

$$X \sim f(x) \Rightarrow \alpha + \beta X \sim \frac{1}{\beta} f(\alpha + \beta x) \tag{58}$$

6 Joint Distributions and Independence

bivariate distribution is **independent** if

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \tag{59}$$

a multivariate distribution is **independent** if

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i}(x_i) \tag{60}$$

a multivariate distribution can be factored into **marginal** and **conditional** distributions

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)\, f_Y(y) \Rightarrow f_{X|Y} = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{61}$$

if the multivariate distribution is **independent identically distributed (IID)**

$$X_1, \ldots, X_n \sim X \Rightarrow f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_X(x_i) \tag{62}$$

7 Common Functions

- Error function and complementary error function

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt \tag{63}$$

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) \tag{64}$$

- Gamma function

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \tag{65}$$

$$\begin{aligned}
\Gamma(x) &= \int_0^\infty t^{x-1} e^{-t}\, dt \\
&= -t^{x-1} e^{-t}\Big|_0^\infty - \int_0^\infty (x-1) t^{x-2} (-e^{-t})\, dt \\
&= (x-1) \int_0^\infty t^{x-2} e^{-t}\, dt \\
&= (x-1)\Gamma(x-1)
\end{aligned} \tag{66}$$

$$\Gamma(n) = (n-1)! \tag{67}$$

- Beta function

$$\mathrm{B}(x,y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt \tag{68}$$

$$\begin{aligned}
\Gamma(x)\Gamma(y) &= \int_0^\infty s^{x-1} e^{-s}\, ds \int_0^\infty t^{y-1} e^{-t}\, dt \\
&= \int_0^\infty \int_0^\infty s^{x-1} t^{y-1} e^{-(s+t)}\, ds\, dt && \begin{cases} s = uv \\ t = u(1-v) \end{cases} \\
&= \int_{u=0}^\infty \int_{v=0}^1 (uv)^{x-1} (u(1-v))^{y-1} e^{-u} u\, du\, dv && |J| = u \\
&= \int_0^\infty e^{-u} u^{x+y-1}\, du \int_0^1 v^{x-1} (1-v)^{y-1}\, dv \\
&= \Gamma(x+y)\, \mathrm{B}(x,y)
\end{aligned} \tag{69}$$

$$\mathrm{B}(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \tag{70}$$

$$\begin{aligned} \mathrm{B}\left(\frac{1}{2},\frac{1}{2}\right) &= \int_0^1 t^{-\frac{1}{2}}(1-t)^{-\frac{1}{2}}\,dt \\ &= 2\int_0^{\frac{\pi}{2}} \frac{\cos\theta\sin\theta}{\cos\theta\sin\theta}\,d\theta \\ &= \pi \end{aligned} \tag{71}$$

$$\mathrm{B}\left(\frac{1}{2},\frac{1}{2}\right) = \frac{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(1)} \Rightarrow \Gamma\left(\frac{1}{2}\right) = \sqrt{\mathrm{B}\left(\frac{1}{2},\frac{1}{2}\right)} = \sqrt{\pi} \tag{72}$$

- Multivariate Beta Function

$$\mathrm{B}(\alpha_1,\ldots,\alpha_n) = \frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^n \alpha_i\right)} \tag{73}$$

$$\mathrm{B}(\alpha_1,\ldots,\alpha_n) = \frac{\Gamma(\alpha_j)\prod_{i\neq j}\Gamma(\alpha_i)}{\Gamma\left(\alpha_j + \sum_{i\neq j}\alpha_i\right)} = \frac{\prod_{i\neq j}\Gamma(\alpha_i)}{\Gamma(\sum_{i\neq j}\alpha_i)}\,\mathrm{B}\left(\alpha_j, \sum_{i\neq j}\alpha_i\right) \tag{74}$$

8 Common Distributions

8.1 Discrete Distributions

8.1.1 Sampling With Replacement

8.1.1.1 Bernoulli

$$\mathrm{Ber}(p) \equiv f(k|p) = p^k(1-p)^{1-k}, k \in \{0,1\} \tag{75}$$

8.1.1.2 Binomial

$$X_i \sim \mathrm{Ber}(p) \Rightarrow Y = \sum_{i=1}^n X_i \sim \mathrm{Bin}(n,p) \tag{76}$$

$$\mathrm{Bin}(n,p) \equiv f(k|n,p) = \binom{n}{k} p^k(1-p)^{n-k}, k \in \{0,\cdots,n\} \tag{77}$$

8.1.1.3 Negative Binomial

$$\mathrm{NB}(k|r,p) = \mathrm{Bin}(k|k+r-1,p)\,\mathrm{Ber}(0|p) \tag{78}$$

$$\mathrm{NB}(k|r,p) \equiv f(k|r,p) = \binom{k+r-1}{k} p^k(1-p)^{r-k}, k \in \mathbb{N} \tag{79}$$

8.1.1.4 Geometric

$$\mathrm{Geo}(k|p) = \mathrm{NB}(k|1,1-p) = p(1-p)^{k-1}, k \in \mathbb{N} \tag{80}$$

8.1.1.5 Poisson

$$\mathrm{Poi}(\lambda) = f(k|\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}, k \in \mathbb{N} \tag{81}$$

$$\begin{aligned}
\mathrm{Bin}\left(n, \frac{\lambda}{n}\right) &= f\left(k \middle| n, \frac{\lambda}{n}\right) = \binom{n}{k}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{k}\right)^{n-k} \\
&= \frac{n \cdot n-1 \cdot \cdots \cdot n-k+1}{k!}\frac{\lambda^k}{n^k}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^k \\
&= \left[\left(\frac{n}{n}\right) \cdot \left(\frac{n-1}{n}\right) \cdot \cdots \cdot \left(\frac{n-k+1}{n}\right)\right] \cdot \left[\left(1 - \frac{\lambda}{n}\right)^{-k}\right] \cdot \left[\frac{\lambda^k}{k!}\left(1 - \frac{\lambda}{n}\right)^n\right]
\end{aligned} \tag{82}$$

$$\lim_{n \to \infty} \mathrm{Bin}\left(n, \frac{\lambda}{n}\right) = \frac{\lambda^k}{k!}e^{-\lambda} \equiv \mathrm{Poi}(\lambda) \tag{83}$$

8.1.1.6 Multinomial

The multinomial distribution is realized from the sum of n repeated *dependent* Bernoulli trials, each parametrized by potentially different probabilities of individual success, p_i, and linked by the requirement that one, and only one, may be successful on any given trial, $\sum_{i=1}^k p_i = 1$:

$$\mathrm{Mul}(n, p_1, \cdots, p_k) \equiv f(x_1, \cdots, x_k | n, p_1, \cdots, p_k) = \frac{n}{\prod_{i=1}^k x_i!}\prod_{i=1}^k p_i^{x_i} = \frac{\Gamma(1 + \sum_{i=1}^k x_i)}{\prod_{i=1}^k \Gamma(1 + x_i)}\prod_{i=1}^k p_i^{x_i} \tag{84}$$

$$X_1, \cdots, X_k \sim \mathrm{Mul}(n, p_1, \cdots, p_k) \Rightarrow \begin{cases} \mathbb{E}X_i = np_i \\ \mathbb{V}X_i = np_i(1 - p_i) \\ \mathbb{C}(X_i, X_j) = -np_ip_j, i \neq j \end{cases} \tag{85}$$

8.1.2 Sampling Without Replacement

8.1.2.1 Hypergeometric

$$\mathrm{Hyp}(n, N, K) \equiv f(k|n, N, K) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{86}$$

8.1.2.2 Multivariate Hypergeometric

$$\left.\begin{aligned} \mathbf{k} &= (k_1, \ldots, k_n)^\top \\ \mathbf{K} &= (K_1, \ldots, K_n)^\top \end{aligned}\right\} \Rightarrow \mathrm{MHG}(\mathbf{k}|\mathbf{K}) = \frac{\binom{K_1}{k_1}\cdots\binom{K_n}{k_n}}{\binom{\sum_{i=1}^n K_i}{\sum_{i=1}^n k_i}} \tag{87}$$

8.2 Continuous Distributions

$$f(x|\theta) = h(x)g(\theta)e^{\eta(\theta)T(x)} \tag{88}$$

8.2.1 Gaussian

8.2.1.1 Univariate Gaussian

$$\mathrm{N}(\mu,\sigma^2) \equiv f_{\mathrm{N}}(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{89}$$

$$\mathbb{P}_{\mathrm{N}(\mu,\sigma^2)}[-\infty,x] = F_{\mathrm{N}}(x|\mu,\sigma^2) = \int_{-\infty}^{x} f_{\mathrm{N}}(x|\mu,\sigma^2)\,dx = \frac{1}{2}\left(1+\mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right) \tag{90}$$

$$\mathbb{E}(X-\mu)^{2n-1} = 0 \tag{91}$$

$$g(\alpha) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\alpha\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \frac{1}{\sqrt{\alpha}} \tag{92}$$

$$\mathbb{E}(X-\mu)^{2n} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^{2n} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = (-2\sigma^2)^n \left.\frac{d^n}{d\alpha^n} g(\alpha)\right|_{\alpha=1} = (2n-1)!!\,\sigma^{2n} \tag{93}$$

$$\begin{aligned} M_{\mathrm{N}(\mu,\sigma^2)}(t) \equiv \mathbb{E}e^{tX} &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2-tx}{2\sigma^2}}\,dx && (94) \\ &= e^{t\mu+\frac{1}{2}t^2\sigma^2} \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-(\mu+t\sigma^2))^2}{2\sigma^2}}\,dx && (95) \\ &= e^{t\mu+\frac{1}{2}t^2\sigma^2} && (96) \end{aligned}$$

$$\phi_{\mathrm{N}(\mu,\sigma^2)}(t) \equiv \mathbb{E}e^{itX} = e^{it\mu-\frac{1}{2}t^2\sigma^2} \tag{97}$$

$$\left.\begin{array}{l} X \sim \mathrm{N}(\mu,\sigma^2) \Rightarrow M_X = e^{t\mu+\frac{1}{2}t^2\sigma^2} \\ Y \sim \mathrm{N}(\nu,\tau^2) \Rightarrow M_Y = e^{t\nu+\frac{1}{2}t^2\tau^2} \end{array}\right\} \Rightarrow M_{X+Y} = e^{t(\mu+\nu)+\frac{1}{2}t^2(\sigma^2+\tau^2)} \Rightarrow X+Y \sim \mathrm{N}(\mu+\nu,\sigma^2+\tau^2) \tag{98}$$

8.2.1.2 Standard Normal

$$Z \sim \mathrm{N}(0,1) \tag{99}$$

8.2.1.3 Multivariate Gaussian

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \Rightarrow \begin{cases} \boldsymbol{\mu} = \begin{pmatrix} \mathbb{E}X_1 \\ \vdots \\ \mathbb{E}X_n \end{pmatrix} \\ \boldsymbol{\Sigma} = \begin{pmatrix} \mathbb{V}X_1 & \cdots & \mathbb{C}(X_1,X_n) \\ \vdots & \ddots & \vdots \\ \mathbb{C}(X_1,X_n) & \cdots & \mathbb{V}X_n \end{pmatrix} \end{cases} \tag{100}$$

$$\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \det\boldsymbol{\Sigma}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{101}$$

$$\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\mathbf{z}^\top \mathbf{z}} \tag{102}$$

Transformation, $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{Z}$,

$$\left.\begin{aligned} &\mathbb{E}\mathbf{X} \equiv \mathbb{E}(\boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{Z}) = \boldsymbol{\mu} \\ &\mathbb{V}\mathbf{X} \equiv \mathbb{E}\left(\mathbf{X} - \mathbb{E}\mathbf{X}\right)\left(\mathbf{X} - \mathbb{E}\mathbf{X}\right)^\top = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top \end{aligned}\right\} \Rightarrow \mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top) \tag{103}$$

$$\boldsymbol{\Sigma} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top \Rightarrow \boldsymbol{\Gamma} = \mathbf{Q}\mathbf{D}^{\frac{1}{2}}, \qquad \mathbf{D}^{\frac{1}{2}} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n) \tag{104}$$

Transformation, $\mathbf{Z} = \boldsymbol{\Gamma}^{-1}(\mathbf{X} - \boldsymbol{\mu})$

$$\left.\begin{aligned} &\mathbb{E}\mathbf{Z} \equiv \mathbb{E}\boldsymbol{\Gamma}(\mathbf{X} - \boldsymbol{\mu}) = \mathbf{0} \\ &\mathbb{V}\mathbf{Z} \equiv \mathbb{E}\left(\mathbf{Z} - \mathbb{E}\mathbf{Z}\right)\left(\mathbf{Z} - \mathbb{E}\mathbf{Z}\right)^\top = \mathbb{E}\mathbf{Z}\mathbf{Z}^\top = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-\top} = \mathbf{I} \end{aligned}\right\} \Rightarrow \mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}) \tag{105}$$

$$\boldsymbol{\Sigma} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top \Rightarrow \boldsymbol{\Sigma}^{-1} = \mathbf{Q}\mathbf{D}^{-1}\mathbf{Q}^\top \tag{106}$$

$$\mathbf{D} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2) \Rightarrow \mathbf{D}^{-1} = \mathrm{diag}\left(\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_n^2}\right) \tag{107}$$

$$\mathbf{y} = \mathbf{Q}^\top(\mathbf{x} - \boldsymbol{\mu}) \Rightarrow \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \frac{1}{2}\mathbf{y}^\top \mathbf{D}^{-1}\mathbf{y} = \sum_{i=1}^{n} \frac{y_i^2}{2\sigma_i^2} \tag{108}$$

$$\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}} e^{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})} = \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{y_i^2}{2\sigma_i^2}} = \prod_{i=1}^{n} f_{\mathrm{N}}(y_i | 0, \sigma_i^2) \tag{109}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \Rightarrow \begin{cases} \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \\ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \end{cases} \tag{110}$$

$$\mathbf{X}_1 | (\mathbf{X}_2 = \mathbf{x}_2) \sim \mathrm{N}(\boldsymbol{x}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \tag{111}$$

8.2.1.4 Marginal and Conditional Gaussian Distributions $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \Rightarrow \begin{cases} \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \\ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \end{cases} \tag{112}$$

$$(\mathbf{x}^\top - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}\,(\mathbf{x}^\top - \boldsymbol{\mu}) = \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu} \\ \mathbf{x}_2 - \boldsymbol{\mu} \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu} \\ \mathbf{x}_2 - \boldsymbol{\mu} \end{pmatrix} \tag{113}$$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} I_p & 0 \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & I_q \end{pmatrix} \begin{pmatrix} \left(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)^{-1} & 0 \\ 0 & \boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix} \begin{pmatrix} I_p & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ 0 & I_q \end{pmatrix} \tag{114}$$

$$\begin{aligned}&(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\\ =&\begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}\\ \mathbf{x}_2-\boldsymbol{\mu}\end{pmatrix}^\top \begin{pmatrix}I_p\\ \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\end{pmatrix}\left(\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)^{-1}\begin{pmatrix}I_p\\ -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\end{pmatrix}^\top \begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}\\ \mathbf{x}_2-\boldsymbol{\mu}\end{pmatrix}+(\mathbf{x}_2-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\end{aligned} \tag{115}$$

$$\mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})=\mathrm{N}(\boldsymbol{\mu}_1-(\mathbf{x}_2-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21},\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})\ \mathrm{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_{22}) \tag{116}$$

$$\text{Marginal Distribution: }\mathbf{X}_2\sim \mathrm{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_{22}) \tag{117}$$

$$\text{Conditional Distribution: }\mathbf{X}_1|(\mathbf{X}_2=\mathbf{x}_2)\sim \mathrm{N}(\boldsymbol{\mu}_1-(\mathbf{x}_2-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21},\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}) \tag{118}$$

8.2.1.5 Mean and Variance of IID Normal Random Variables

$$\left.\begin{array}{l}\mathbf{X}=(X_1,\cdots,X_n)^\top, X_i\sim \mathrm{N}(\mu,\sigma^2)\\ \mathbf{1}_n=(1,\cdots,1)^\top\end{array}\right\}\Rightarrow \mathbf{X}\sim \mathrm{N}(\mu\mathbf{1}_n,\sigma^2 I_n) \tag{119}$$

let the random vector, $\mathbf{Y}$, be formed by the linear transformaition of $\mathbf{X}$ by the $n\times n$ orthogonal matrix, Q

$$\mathbf{Y}=(\mathbf{Y}_1,\cdots,\mathbf{Y}_n)^\top=Q^\top\mathbf{X}\Rightarrow \mathbf{Y}\sim \mathrm{N}(\mu\mathbf{1}_n,\sigma^2 Q^\top I_n Q)=\mathrm{N}(\mu\mathbf{1}_n,\sigma^2 I_n) \tag{120}$$

Fisher's Theorem: sample mean and sample variance taken from IID normal distribution are independent:

$$\hat\mu\mathbf{1}_n=\mathbf{1}_n\frac{1}{n}\mathbf{1}_n^\top\mathbf{x}=\mathbf{1}_n(\mathbf{1}_n^\top\mathbf{1}_n)^{-1}\mathbf{1}_n^\top\mathbf{x}=P_{\mathbf{1}_n}\mathbf{x} \tag{121}$$

$$\hat\sigma^2=\frac{1}{n-1}(\mathbf{x}-\hat\mu\mathbf{1}_n)^\top(\mathbf{x}-\hat\mu\mathbf{1}_n)=\frac{1}{n-1}(\mathbf{x}-P_{\mathbf{1}_n}\mathbf{x})^\top(\mathbf{x}-P_{\mathbf{1}_n}\mathbf{x})=\frac{1}{n-1}\mathbf{x}^\top(I_n-P_{\mathbf{1}_n})\mathbf{x} \tag{122}$$

These equations imply that, for arbitrary sample sets of data, information on the mean and variances of the distributed data are carried in mutually orthogonal 1- and $(n-1)$-dimensional subspaces, respectively. Knowledge of the mean carries no information about the variance, and *v.v.*

Sums of Gaussian are Gaussian

$$\sqrt{n}\hat\mu\sim \mathrm{N}(\mu,\sigma^2) \tag{123}$$

Furthermore, the distribution of the sample variance can be shown to be chi-square distributed:

$$(n-1)\frac{\hat\sigma^2}{\sigma^2}=\frac{1}{\sigma^2}\mathbf{x}^\top(I_n-P_{\mathbf{1}_n})\mathbf{x}=\sum_{i=2}^{n}\frac{y_i^2}{\sigma^2}=\sum_{i=2}^{n}z_i^2\sim\chi^2_{n-1} \tag{124}$$

8.2.2 Gamma-Derived Distributions

8.2.2.1 Gamma

$$\Gamma(\alpha,\beta)\equiv f_\Gamma(x|\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x},\ x\geq 0 \tag{125}$$

$$\begin{aligned}\Gamma(\alpha) &= \int_0^\infty t^{\alpha-1}e^{-t}\,dt \\ &= \int_0^\infty (\beta x)^{\alpha-1}e^{-\beta x}\beta\,dx \qquad t=\beta x \\ &= \int_0^\infty \beta^\alpha x^{\alpha-1}e^{-\beta x}\,dx \end{aligned} \tag{126}$$

$$\begin{aligned}\mathbb{E}X^n &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty x^k x^{\alpha-1}e^{-\beta x}\,dx \\ &= \frac{\beta^\alpha}{\beta^{\alpha+k}}\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}\int_0^\infty \frac{\beta^{\alpha+k}}{\Gamma(\alpha+k)}x^{\alpha+k-1}e^{-\beta x}\,dx \\ &= \frac{1}{\beta^k}\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}\int_0^\infty f_\Gamma(x|\alpha+k,\beta)\,dx \\ &= \frac{1}{\beta^k}\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \end{aligned} \tag{127}$$

$$\mathbb{E}X = \frac{\alpha}{\beta} \tag{128}$$

$$\mathbb{V}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{(\alpha+1)\alpha}{\beta^2} - \frac{\alpha^2}{\beta^2} = \frac{\alpha}{\beta^2} \tag{129}$$

$$\begin{aligned}M_{\Gamma(\alpha,\beta)} \equiv \mathbb{E}e^{tX} &= \int_0^\infty e^{tx}\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-(\beta-t)x}\,dx \\ &= \frac{\beta^\alpha}{(\beta-t)^\alpha}\int_0^\infty \frac{(\beta-t)^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-(\beta-t)x}\,dx \\ &= \left(\frac{\beta}{\beta-t}\right)^\alpha \int_0^\infty f_\Gamma(x|\alpha,\beta-t)\,dx \\ &= \left(\frac{\beta}{\beta-t}\right)^\alpha \end{aligned} \tag{130}$$

$$X_i \sim \Gamma(\alpha_i,\beta) \Rightarrow \sum_{i=1}^n X_i \sim \Gamma\left(\sum_{i=1}^n \alpha_i, \beta\right) \tag{131}$$

8.2.2.2 Chi-square

$$X \sim \mathrm{N}(0,1) \Rightarrow X^2 \sim \chi^2 = \Gamma\left(\frac{1}{2},\frac{1}{2}\right) \tag{132}$$

$$\mathbb{P}\{X^2 \le x\} = \mathbb{P}\{-\sqrt{x} \le X \le \sqrt{x}\} = \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\,dt \tag{133}$$

$$
\begin{aligned}
\chi^2 \equiv f_{\chi^2}(x) &= \frac{d}{dx}\int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt \\
&= \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x}{2}}\right)\left(\frac{1}{2}x^{-\frac{1}{2}}\right) - \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x}{2}}\right)\left(-\frac{1}{2}x^{-\frac{1}{2}}\right) \\
&= \frac{1}{\sqrt{2\pi}} x^{\frac{1}{2}-1} e^{-\frac{x}{2}} \\
&= \Gamma\left(\frac{1}{2}, \frac{1}{2}\right) \qquad (134)
\end{aligned}
$$

$$X_i \sim \mathrm{N}(0,1) \Rightarrow \sum_{i=1}^{n} X_i^2 \sim \chi_n^2 = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right) \qquad (135)$$

$$X \sim \chi_n^2 \Rightarrow \begin{cases} \mathbb{E}X = n \\ \mathbb{V}X = \frac{n}{2} \end{cases} \qquad (136)$$

$$\mathbf{\Sigma} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top = \mathbf{\Gamma}\mathbf{\Gamma}^\top \Rightarrow \mathbf{Z} = \mathbf{\Gamma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \qquad (137)$$

$$\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{\Sigma}) \Rightarrow (\mathbf{X} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) = \mathbf{Z}^\top \mathbf{Z} \sim \chi_n^2 \qquad (138)$$

8.2.2.3 Inverse Gamma

8.2.3 Beta-Derived Distributions

8.2.3.1 Beta

$$\mathrm{B}(\alpha,\beta) \equiv f_{\mathrm{B}}(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 \le x \le 1 \qquad (139)$$

$$
\begin{aligned}
\mathbb{E}X^k &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^k x^{\alpha-1}(1-x)^{\beta-1}\,dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+k)\Gamma(\beta)}{\Gamma(\alpha+\beta+k)} \int_0^1 \frac{\Gamma(\alpha+\beta+k)}{\Gamma(\alpha+k)\Gamma(\beta)} x^{\alpha+k-1}(1-x)^{\beta-1}\,dx \\
&= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+k)} \int_0^1 f_{\mathrm{B}}(x|\alpha+k,\beta)\,dx \\
&= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+k)} \qquad (140)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}X &= \frac{\alpha}{\alpha+\beta} \qquad (141) \\
\mathbb{V}X &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2} = \frac{\alpha\beta}{(\alpha+\beta)^2} \qquad (142)
\end{aligned}
$$

8.2.3.2 Dirichlet

$$\mathrm{Dir}(\alpha_1,\cdots,\alpha_n) \equiv f_{\mathrm{D}}(x_1,\cdots,x_n|\alpha_1,\cdots,\alpha_n) = \frac{\prod_{i=1}^{n} x_i^{\alpha_i-1}}{\mathrm{B}(\alpha_1,\cdots,\alpha_n)}, \begin{cases} 0 \le x_i \le 1 \\ \sum_{i=1}^{n} x_i = 1 \end{cases} \tag{143}$$

marginal distributions

$$f_D(x_i|\alpha_1,\cdots,\alpha_n) = f_{\mathrm{B}}\left(x_i \middle| \alpha_i, \sum_{j\neq i}\alpha_j\right) = \frac{x_i^{\alpha_i-1}(1-x_i)^{\sum_{j\neq i}\alpha_j-1}}{\mathrm{B}(\alpha_i,\sum_{j\neq i}\alpha_j)} \tag{144}$$

$$\mathbb{E}X_i = \frac{\alpha_i}{\sum_{j=1}^{n}\alpha_j} \tag{145}$$

$$\mathbb{V}X_i = \frac{\alpha_i\sum_{j\neq i}\alpha_j}{\left(\sum_{j=1}^{n}\alpha_j\right)^2} \tag{146}$$

8.2.4 Distributions of Ratios of Standard Normal Random Variables

8.2.4.1 F-Distribution

$$\left.\begin{matrix} U_1,\ldots,U_k \\ V_1,\ldots,V_m \end{matrix}\right\} \sim Z = \mathrm{N}(0,1) \Rightarrow \frac{\frac{1}{k}\sum_{i=1}^{k}U_i^2}{\frac{1}{m}\sum_{j=1}^{m}V_j^2} \sim F(k,m),$$
$$F(k,m) \equiv f_F(x|k,m) = \frac{\Gamma(\frac{k+m}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}\left(\frac{k}{m}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1}\left(1+\frac{k}{m}x\right)^{-\frac{k+m}{2}} \tag{147}$$

Define the numerator and denominator in terms of chi-square distributed variables

$$\left.\begin{matrix} U = \sum_{i=1}^{k}U_i^2 \sim \chi_k^2 \\ V = \sum_{j=1}^{m}V_j^2 \sim \chi_m^2 \end{matrix}\right\} \Rightarrow \mathbb{P}\left\{\frac{\frac{U}{k}}{\frac{V}{m}} \le x\right\} = \mathbb{P}\{U \le \frac{k}{m}xV\}$$
$$= \iint_{\{U\le\frac{k}{m}xV\}} f_{\chi_k^2}(u)f_{\chi_m^2}(v)\,du\,dv = \int_0^\infty\int_0^{\frac{k}{m}xv} f_{\chi_k^2}(u)f_{\chi_m^2}(v)\,du\,dv \tag{148}$$

$$\begin{aligned}
F(k,m) \equiv f_F(x|k,m) &= \frac{d}{dx}\int_0^{\infty}\int_0^{\frac{k}{m}xv} f_{\chi^2_k}(u) f_{\chi^2_m}(v)\,du\,dv = \int_0^{\infty}\frac{d}{dx}\left(\int_0^{\frac{k}{m}xv} f_{\chi^2_k}(u)\,du\right) f_{\chi^2_m}(v)\,dv \\
&= \frac{k}{m}\int_0^{\infty} f_{\chi^2_k}\left(\frac{k}{m}xv\right) f_{\chi^2_m}(v)\,v\,dv \\
&= \frac{k}{m}\int_0^{\infty}\left(\frac{\frac{1}{2}^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)}\left(\frac{k}{m}xv\right)^{\frac{k}{2}-1} e^{-\frac{\frac{k}{m}xv}{2}}\right)\left(\frac{\frac{1}{2}^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)} v^{\frac{m}{2}-1} e^{-\frac{v}{2}}\right) v\,dv \\
&= \frac{k^{\frac{k}{2}}}{m}\int_0^{\infty}\frac{\left(\frac{1}{2}\right)^{\frac{k+m}{2}} x^{\frac{k}{2}-1}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} v^{\frac{k+m}{2}-1} e^{-\frac{1}{2}v\left(\frac{k}{m}t+1\right)}\,dv \\
&= \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)}\left(\frac{k}{m}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1}\left(\frac{1}{\frac{k}{m}x+1}\right)^{\frac{k+m}{2}}\int_0^{\infty}\frac{\left(\frac{t+1}{2}\right)^{\frac{k+m}{2}}}{\Gamma\left(\frac{k+m}{2}\right)} v^{\frac{k+m}{2}-1} e^{-v\frac{t+1}{2}}\,dv \\
&= \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)}\left(\frac{k}{m}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1}\left(1+\frac{k}{m}x\right)^{-\frac{k+m}{2}}\int_0^{\infty} f_{\Gamma}\left(v\,\middle|\,\frac{k+m}{2},\frac{t+1}{2}\right)dv \\
&= \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)}\left(\frac{k}{m}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1}\left(1+\frac{k}{m}x\right)^{-\frac{k+m}{2}} \qquad (149)
\end{aligned}$$

8.2.4.2 T-Distribution

$$\left.\begin{matrix} U_1 \\ V_1,\dots,V_m \end{matrix}\right\} \sim Z = \mathrm{N}(0,1) \Rightarrow \frac{U_1}{\sqrt{\frac{1}{m}\sum_{j=1}^{m} V_j^2}} \sim T(m),$$

$$T(m) \equiv f_T(x|m) = \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{m}{2})}\frac{1}{\sqrt{m}}\left(1+\frac{x^2}{m}\right)^{-\frac{m+1}{2}} \qquad (150)$$

$$\left.\begin{matrix} U = U_i^2 \sim \chi^2 \\ V = \sum_{j=1}^{m} V_j^2 \sim \chi^2_m \end{matrix}\right\} \;\Rightarrow\; \mathbb{P}\left\{-x \le \frac{\sqrt{U}}{\sqrt{\frac{1}{m}V}} \le x\right\} \;=\; \mathbb{P}\left\{\frac{U}{\frac{1}{m}V} \le x^2\right\} \;=\; \int_0^{x^2} f_F(v|1,m)\,dv \qquad (151)$$

$$\begin{aligned}
T(m) \equiv f_T(x|m) &= \frac{1}{2}\frac{d}{dx}\int_0^{x^2} f_F(v|1,m)\,dv = x f_F\left(x^2|1,m\right) \\
&= \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{m}{2}\right)}\frac{1}{\sqrt{m}}\left(1+\frac{x^2}{m}\right)^{-\frac{m+1}{2}} \qquad (152)
\end{aligned}$$

8.2.4.3 Cauchy

$$\left.\begin{matrix} U \\ V \end{matrix}\right\} \sim Z = \mathrm{N}(0,1) \Rightarrow \frac{U}{V} \sim \mathrm{Cau}(0,1) \equiv f_C(x) = \frac{1}{\pi}\frac{1}{x^2+1} \qquad (153)$$

$$\begin{aligned}
\mathbb{P}\left\{\frac{U}{V} \le x\right\} = \mathbb{P}\left\{U \le xV\right\} &= \iint_{\{u \le xv\}} f_{\mathrm{N}}(u|0,1) f_{\mathrm{N}}(v|0,1)\,du\,dv \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{xv} f_{\mathrm{N}}(u|0,1) f_{\mathrm{N}}(v|0,1)\,du\,dv \qquad (154)
\end{aligned}$$

$$
\begin{aligned}
\mathrm{Cau}(0,1) \equiv f_{\mathrm{C}}(x) &= \frac{d}{dx}\mathbb{P}\left\{\frac{U}{V} \le x\right\} = \frac{d}{dx}\int_0^\infty \int_0^{xv} f_{\mathrm{N}}(u|0,1) f_{\mathrm{N}}(v|0,1)\, du\, dv \\
&= \int_{-\infty}^\infty \left(\frac{d}{dx}\int_{-\infty}^{xv} f_{\mathrm{N}}(u|0,1)\, du\right) f_{\mathrm{N}}(v|0,1)\, dv = \int_{-\infty}^\infty f_{\mathrm{N}}(xv|0,1) f_{\mathrm{N}}(v|0,1)\, dv \\
&= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2v^2}{2}} \frac{1}{2\pi} e^{-\frac{v^2}{2}} v\, dv = \frac{1}{2\pi}\int_{-\infty}^\infty e^{-\frac{v^2(x^2+1)}{2}} v\, dv \\
&= \frac{1}{2\pi}\frac{1}{x^2+1}\int_{-\infty}^\infty e^{-t}\, dt = \frac{1}{\pi}\frac{1}{x^2+1}\int_0^\infty e^{-t}\, dt \\
&= \frac{1}{\pi}\frac{1}{x^2+1} \qquad (155)
\end{aligned}
$$

$$\mathrm{Cau}(0,1) \equiv T(1) \qquad (156)$$

8.2.5 Other Common Distributions

8.2.5.1 Exponential

$$\mathrm{Exp}(\lambda) \equiv f_{\mathrm{E}}(x|\lambda) = \lambda e^{-\lambda x}, \qquad x \ge 0 \qquad (157)$$

memoryless:

$$\mathbb{P}\{x > s + t | x > s\} = \mathbb{P}\{x > t\} \qquad (158)$$

$$h(t) = \frac{f(t)}{1 - \int_0^t f(x)\, dx} = \lambda \qquad (159)$$

8.2.5.2 Pareto

$$\mathrm{Par}(\alpha, x_m) \equiv f_{\mathrm{P}}(x|\alpha, x_m) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, & x \ge x_m \\ 0, & x < x_m \end{cases} \qquad (160)$$

hazard rate (burn-in period)

$$h(t) = \frac{\alpha}{t} \qquad (161)$$

8.2.5.3 Weibull

8.2.5.4 Uniform

$$\mathrm{Uni}(0,1) \equiv f_{\mathrm{U}}(x) = 1, \;\; 0 \le x \le 1 \qquad (162)$$

9 Order Statistics

Order statistics are the probability distributions of **cumulative probability rank**, or ‘percentile’, of finite samples taken with replacement from arbitrary distributions.

Let a set of n IID random variables, designated as $X_1, \cdots, X_n$, be sampled from a uniform distribution, $X_i \sim \mathrm{Uni}(0,1)$. The random variables sorted in increasing order, designated as $X_{(1)}, \cdots, X_{(n)}$,

- $k-1$ events fall within $[0,u)$
- 1 event falls within $[u+du)$
- $n-k$ events fall within $[u+du,1]$

finite-sized intervals given by multinomial probabilities,

$$\frac{n!}{(k-1)!1!(n-k)!}u^{k-1}\cdot du\cdot(1-u-du)^{n-k} \tag{163}$$

$$X_{(k)}\sim \mathrm{B}(k,n+1-k) \tag{164}$$

10 Asymptotic Limits

10.1 Convergence of Random Variables

10.1.1 Convergence in Distribution

$$\lim_{n\to\infty}\mathbb{P}_{X_n}\{X_n<x\}=\mathbb{P}_X\{X<x\}\Leftrightarrow X_n\xrightarrow{d}X \tag{165}$$

10.1.2 Convergence in Probability (Weak Convergence)

$$\lim_{n\to\infty}\mathbb{P}\{|X_n-X|\geq\epsilon\}=0\Leftrightarrow X_n\xrightarrow{p}X \tag{166}$$

10.1.3 Convergence Almost Surely (Strong Convergence)

$$\mathbb{P}\{\lim_{n\to\infty}|X_n-X|\leq\epsilon\}=0\Leftrightarrow X_n\xrightarrow{a.s.}X \tag{167}$$

10.1.4 Product of Convergent Random Variables

Slutsky's Theorem

$$\left.\begin{matrix}X_n\xrightarrow{d}X\\ Y_n\xrightarrow{p}c\end{matrix}\right\}\Rightarrow X_nY_n\xrightarrow{d}cX \tag{168}$$

10.2 Asymptotic Limits of IID Samples

10.2.1 Law of Large Numbers

$$\mathbb{E}X=\mu\Rightarrow\mathbb{E}\bar{X}_n=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i=\mu \tag{169}$$

Weak Law of Large Numbers

$$\lim_{n\to\infty}\mathbb{P}\{|\bar{X}_n-\mu|\geq\epsilon\}=0 \tag{170}$$

Strong Law of Large Numbers

$$\mathbb{P}\left\{\lim_{n\to\infty} |\bar{X}_n - \mu| \geq \epsilon\right\} = 0 \tag{171}$$

10.2.2 Central Limit Theorem

10.2.2.1 Univariate Theorem

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathrm{N}(0,1) \tag{172}$$

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \Rightarrow \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma\sqrt{n}} \tag{173}$$

$$\left.\begin{array}{l} \mathbb{E}\left(\frac{X_i-\mu}{\sigma\sqrt{n}}\right) = 0 \\ \mathbb{E}\left(\frac{X_i-\mu}{\sigma\sqrt{n}}\right)^2 = \frac{1}{n} \\ \phi_X(t) = 1 + it\mathbb{E}X - \frac{t^2}{2}\mathbb{E}X^2 + \cdots \end{array}\right\} \Rightarrow \phi_{\frac{X_i-\mu}{\sigma\sqrt{n}}}(t) = 1 - \frac{t^2}{2n} + \mathcal{O}\left(n^{-\frac{3}{2}}\right) \tag{174}$$

$$\phi_{\sum_{i=1}^{n} \frac{X_i-\mu}{\sigma\sqrt{n}}}(t) = \prod_{i=1}^{n} \phi_{\frac{X_i-\mu}{\sigma\sqrt{n}}}(t) = \prod_{i=1}^{n}\left(1 - \frac{t^2}{2n} + \mathcal{O}\left(n^{-\frac{3}{2}}\right)\right) \tag{175}$$

$$\lim_{n\to\infty}\prod_{i=1}^{n}\left(1 - \frac{t^2}{n} + \mathcal{O}\left(n^{-\frac{3}{2}}\right)\right) = \lim_{n\to\infty}\left(1 - \frac{t^2}{n} + \mathcal{O}\left(n^{-\frac{3}{2}}\right)\right)^n = e^{-\frac{t^2}{2}} = \phi_{\mathrm{N}(0,1)}(t) \tag{176}$$

10.2.2.2 Multivariate Theorem

$$\mathbb{E}\mathbf{X} = \boldsymbol{\mu} \Rightarrow \mathbb{E}\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\mathbf{X}_i = \boldsymbol{\mu} \tag{177}$$

$$\mathbb{V}\mathbf{X} = \mathbb{E}\mathbf{X}\mathbf{X}^\top - \mathbb{E}\mathbf{X}\,\mathbb{E}\mathbf{X}^\top = \boldsymbol{\Sigma} \tag{178}$$

$$\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{179}$$

Delta Method

Transformation $g(\mathbf{X})$

$$\left.\begin{array}{l} \mathbb{E}g(\mathbf{X}) = g(\boldsymbol{\mu}) \\ \mathbb{V}g(\mathbf{X}) = \nabla g(\boldsymbol{\mu})^\top \boldsymbol{\Sigma} \nabla g(\boldsymbol{\mu}) \end{array}\right\} \Rightarrow \sqrt{n}\left(g(\mathbf{X}_n) - g(\boldsymbol{\mu})\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \nabla g(\boldsymbol{\mu})^\top \boldsymbol{\Sigma} \nabla g(\boldsymbol{\mu})\right) \tag{180}$$

$g(\mathbf{X}) = \Gamma\mathbf{X}$

$$\left.\begin{array}{l} \mathbb{E}g(\mathbf{X}) = \Gamma\boldsymbol{\mu} \\ \mathbb{V}g(\mathbf{X}) = \Gamma^\top \boldsymbol{\Sigma} \Gamma \end{array}\right\} \Rightarrow \sqrt{n}\Gamma\left(\mathbf{X}_n - \boldsymbol{\mu}\right)) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \Gamma^\top \boldsymbol{\Sigma} \Gamma\right)) \tag{181}$$

11 Likelihood Function and Information Measures

$$\mathbb{I}_S X = -\ln \mathbb{P}_X \tag{182}$$

11.1 Shannon Information and Entropy

$$\mathbb{H}X = \mathbb{E}\mathbb{I}_S X \tag{183}$$

Relative Entropy or the **Kullback-Leibler Divergence**

$$D_{KL}(P||Q) = -\int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)}\, dx = \mathbb{H}X - \mathbb{E}_X \mathbb{I}_S Y \tag{184}$$

by Jensen's Inequality the relative entropy is always positive:

$$D_{KL}(P||Q) = -\int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)}\, dx \geq -\ln \left(\int_{-\infty}^{\infty} p(x) \frac{p(x)}{q(x)}\, dx \right) = -\ln \int_{-\infty}^{\infty} q(x)\, dx = 0 \tag{185}$$

Conditional Entropy

$$\mathbb{H}(X|Y) = \mathbb{H}(X,Y) - \mathbb{H}Y \tag{186}$$

Mutual Information

$$\mathbb{I}_M(X,Y) = \mathbb{H}X + \mathbb{H}Y - \mathbb{H}(X,Y) = D_{KL}\left(p(x,y)||\, p(x)p(y)\right) = \mathbb{E}_Y D_{KL}\left(p(x|y)||\, p(x)\right) \tag{187}$$

11.2 Likelihood Function and Fisher Information

Let $f(\mathbf{x}|\theta)$ be the joint distribution of the sample, $\mathbf{X} = (X_1, \cdots, X_n)$, taken from a distribution parametrized by θ. The **likelihood function**, $L(\theta|\mathbf{x})$, given that $\mathbf{X} = \mathbf{x}$ is observed, is defined as

$$L(\theta|\mathbf{x}) \equiv f(\mathbf{x}|\theta) \tag{188}$$

The **log-likelihood function** is identical to the Shannon information for parametrized distributions

$$\ln L(\theta|\mathbf{x}) = \ln f(\mathbf{x}|\theta) = \ln \mathbb{P}\mathbf{X}_\theta = -\mathbb{I}_S \mathbf{X}_\theta \tag{189}$$

$$L(\theta|\mathbf{x}) \equiv f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \Rightarrow \ln L(\theta|\mathbf{x}) = \sum_{i=1}^{n} \ln f(x_i|\theta) \tag{190}$$

11.2.1 Fisher Information

The **score function**, $S(\theta|\mathbf{x})$, measures the sensitivity of the likelihood function to changes in the parameter value,

$$S(\theta|\mathbf{x}) = \frac{\partial}{\partial\theta} \ln L(\theta|\mathbf{x}) = \frac{\partial}{\partial\theta} \ln f(\mathbf{x}|\theta) = \frac{1}{f(\mathbf{x}|\theta)} \frac{\partial}{\partial\theta} f(\mathbf{x}|\theta) \tag{191}$$

Two identities

$$\int_D f \frac{\partial}{\partial\theta} \ln f \, d\mathbf{x} = \int_D \frac{\partial}{\partial\theta} f \, d\mathbf{x} = \frac{\partial}{\partial\theta} \int_D f \, d\mathbf{x} = 0 \tag{192}$$

$$\int_D f \left(\frac{\partial}{\partial\theta} \ln f \right)^2 d\mathbf{x} = \int_D \frac{1}{f} \left(\frac{\partial f}{\partial\theta} \right)^2 d\mathbf{x} = \int_D \left(\frac{\partial^2 f}{\partial\theta^2} - f \frac{\partial^2}{\partial\theta^2} \ln f \right) d\mathbf{x} = - \int_D f \frac{\partial^2}{\partial\theta^2} \ln f \, d\mathbf{x} \tag{193}$$

The expection of the score function vanishes:

$$\mathbb{E}_\theta S(\theta|\mathbf{X}) = 0 \tag{194}$$

$$\mathbb{V}_\theta S(\theta|\mathbf{X}) = \mathbb{E}_\theta S(\theta|\mathbf{X})^2 - (\mathbb{E}_\theta S(\theta|\mathbf{X}))^2 = \mathbb{E}_\theta S(\theta|\mathbf{X})^2 = -\mathbb{E}_\theta \left(\frac{\partial}{\partial\theta} S(\theta|\mathbf{X}) \right) \tag{195}$$

The **Fisher information** is the variance of the score function,

$$\mathbb{I}_F \mathbf{X}_\theta = \mathbb{V} S(\theta|\mathbf{X}) = -\mathbb{E}_\theta \left(\frac{\partial}{\partial\theta} S(\theta|\mathbf{X}) \right) = -\mathbb{E}_\theta \frac{\partial^2}{\partial\theta^2} \ln f(\mathbf{X}|\theta) \tag{196}$$

Relationship between Shannon entropy and Fisher information:

$$\ln L(\theta + \Delta\theta|\mathbf{x}) \approx \ln L(\theta|\mathbf{x}) + \frac{\partial}{\partial\theta} \ln L(\theta|\mathbf{x}) \Delta\theta + \frac{1}{2} \frac{\partial^2}{\partial\theta^2} \ln L(\theta|\mathbf{x}) \, \Delta\theta^2 \tag{197}$$

$$\begin{aligned} -\mathbb{E}_\theta \ln L(\theta + \Delta\theta|\mathbf{X}) &\approx -\mathbb{E}_\theta \left(\ln L(\theta|\mathbf{X}) + \frac{\partial}{\partial\theta} \ln L(\theta|\mathbf{X}) \Delta\theta + \frac{1}{2} \frac{\partial^2}{\partial\theta^2} \ln L(\theta|\mathbf{X}) \Delta\theta^2 \right) \\ &= \mathbb{E}_\theta \mathbb{I}_S \mathbf{X}_\theta - \mathbb{E}_\theta S(\theta|\mathbf{X}) \Delta\theta + \frac{1}{2} \mathbb{E}_\theta \left(\frac{\partial}{\partial\theta} S(\theta|\mathbf{X}) \right) \Delta\theta^2 \\ \mathbb{H} \mathbf{X}_{\theta + \Delta\theta} &\approx \mathbb{H} \mathbf{X}_\theta + \frac{1}{2} \mathbb{I}_F \mathbf{X}_\theta \Delta\theta^2 \end{aligned} \tag{198}$$

Kullback-Leibler divergence

$$\begin{aligned} D_{KL} \left(L(\theta|\mathbf{x}) || L(\theta + \Delta\theta|\mathbf{x}) \right) &= \int_\Theta L(\theta|\mathbf{x}) \ln \frac{L(\theta|\mathbf{x})}{L(\theta + \Delta\theta|\mathbf{x})} \, d\theta \\ &= \int_\Theta L(\theta|\mathbf{x}) \left(\ln L(\theta|\mathbf{x}) - \ln L(\theta + \Delta\theta|\mathbf{x}) \right) \, d\theta \\ &\approx \int_\Theta L(\theta|\mathbf{x}) \left(-\frac{\partial}{\partial\theta} \ln L(\theta|\mathbf{x}) \Delta\theta - \frac{1}{2} \frac{\partial^2}{\partial\theta^2} \ln L(\theta|\mathbf{x}) \Delta\theta^2 \right) \, d\theta \\ &= -\mathbb{E}_\theta S(\theta|\mathbf{X}) \, \Delta\theta - \frac{1}{2} \mathbb{E}_\theta \frac{\partial}{\partial\theta} S(\theta|\mathbf{X}) \, \Delta\theta^2 \\ &= \frac{1}{2} \mathbb{I}_F \mathbf{X}_\theta \, \Delta\theta^2 \end{aligned} \tag{199}$$

Given a vector of parameters, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^\top$, the **Fisher information matrix** is givenby

$$\mathbb{I}_F \mathbf{X}_{\boldsymbol{\theta}} \equiv -\mathbb{E}_{\boldsymbol{\theta}} \nabla^2 \ln f(\mathbf{X}|\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}} \begin{pmatrix} \frac{\partial^2}{\partial\theta_1^2} \ln f(\mathbf{X}|\boldsymbol{\theta}) & \cdots & \frac{\partial^2}{\partial\theta_1 \partial\theta_m} \ln f(\mathbf{X}|\boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\theta_m \partial\theta_1} \ln f(\mathbf{X}|\boldsymbol{\theta}) & \cdots & \frac{\partial^2}{\partial\theta_m^2} \ln f(\mathbf{X}|\boldsymbol{\theta}) \end{pmatrix} \tag{200}$$

Multidimensional Taylor expansion of the log-likelihood function:

$$\ln L(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}|\mathbf{x}) \approx \ln L(\boldsymbol{\theta}|\mathbf{x}) + \Delta\boldsymbol{\theta}^\top \ln L(\boldsymbol{\theta}|\mathbf{x}) + \frac{1}{2}\Delta\boldsymbol{\theta}^\top H \ln L(\boldsymbol{\theta}|\mathbf{x})\Delta\boldsymbol{\theta} \tag{201}$$

$$D_{KL}\left(L(\boldsymbol{\theta}|\mathbf{x})||L(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}|\mathbf{x})\right) = \Delta\boldsymbol{\theta}^\top \frac{1}{2}\mathbb{I}_F \mathbf{X}_{\boldsymbol{\theta}}\, \Delta\boldsymbol{\theta} \tag{202}$$

12 Bayesian Perspectives

12.1 Bayes' Theorem

Given events, A and B, and a probability measure, $\mathbb{P}$, **Bayes' Theorem** is

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{B|A\}\mathbb{P}\{A\}}{\mathbb{P}\{B\}} \tag{203}$$

12.2 Parameter Refinement and Estimation

the object is to estimate the unknown value of a parameter, θ^*, that governs a parametrized distribution, $f(x|\theta^*)$. The estimate for the parameter is made through a sequence of IID measurements, $x_1, \cdots, x_i$, sampled from the distribution, and each of which adds to the refinement of a distribution for the parameter, $f(\theta)$.

12.3 Conjugate Families

Beta prior, binomial likelihood (n draws):

$$\left.\begin{array}{ll} \text{beta prior:} & \mathrm{B}(\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ \text{binomial likelihood:} & \mathrm{Bin}(n, k) \propto \theta^{k}(1-\theta)^{n-k} \end{array}\right\} \Rightarrow \text{beta posterior: } \mathrm{B}(\alpha + k, \beta + n - k) \tag{204}$$

Gamma prior, Posson likelihood (n draws):

$$\left.\begin{array}{ll} \text{gamma prior:} & \Gamma(\alpha, \beta) \propto x^{\alpha-1}e^{-\beta x} \\ \text{Poisson likelihood:} & \mathrm{Poi}(x) \propto x^{k}e^{-nx} \end{array}\right\} \Rightarrow \text{gamma posterior: } \Gamma(\alpha + k, \beta + n) \tag{205}$$

Gaussian prior, Gaussian likelihood (n draws): This is a model for an iterative determination of the unknown mean of a Gaussian distribution with known variance:

$$\left.\begin{array}{ll} \text{Gaussian prior:} & \mu \sim \mathrm{N}(\mu_0, \sigma_0) \propto \exp\left(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right) \\ \text{Gaussian likelihood:} & \mu|\mathbf{x} \sim \mathrm{N}(\mu, \sigma^2) \propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right) \end{array}\right\}$$

$$\Rightarrow \text{Gaussian posterior: } \mu \sim \mathrm{N}\left(\frac{\kappa_0\mu_0 + n\bar{\mathbf{x}}}{\kappa_0 + n}, \frac{\sigma^2}{\kappa_0 + n}\right) \tag{206}$$

$$\kappa_0 = \frac{\sigma}{\sigma_0} \tag{207}$$

Inverted gamma prior, Gaussian likelihood This is a model for an iterative determination of the unknown variance of a Gaussian distribution with known mean

$$\left.\begin{array}{ll} \text{inverted gamma prior:} & \sigma^2 \sim \mathrm{I}\,\Gamma(\alpha,\beta) \propto \sigma^{-2(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right) \\ \text{Gaussian likelihood:} & \sigma^2|\mathbf{x} \sim \mathrm{N}(\mu,\sigma^2) \propto \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right) \end{array}\right\}$$

$$\Rightarrow \text{inverted gamma posterior: } \sigma^2 \sim \mathrm{I}\,\Gamma\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \tag{208}$$

12.4 Monte Carlo Methods