Probability Notes

Mark DiBattista

February 5, 2020

Abstract

Probability theory is the primary mathematical basis for modeling uncertain physical phenomena and for establishing statistical claims. These notes describe the theoretical setting, a few operational tools, and the common discrete and continuous distributions that form the basic theory, as well as introduce ideas in asymptotic limits, information and entropy, and Bayesian methods that comprise standard, more advanced tools for analysis. Stress is placed on the linear algebraic structure of covariance matrices that partly define multivariate distributions.

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. The acronym, *SN*, refers to the writeup, *Statistics (Notes)*.

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 following notions:

- 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.

The main element covered in the notes below is the mapping function – the random variable – that organizes events in the probability space into the indistiguishable collections assigned weights in the state space. The main mathematical operation in probability is integration, here satisfied by the standard tools of Riemann theory. Only a single exception below requires the finer tools of Lebesgue theory not covered in these notes – the strong convergence of random variables, described below in §9.

The material in this section provides more context than specific tools for calculation, but it is helpful to understand the setting within which random variables are defined.

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,

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

and

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

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, Ω is-$\mathcal{F}$-measurable	$A \subset \Omega$	$\mathcal{F} = \{A_{\lambda\in\Lambda}\}$ (10)
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**,

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

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 Joint Random Variables and Random Vectors

The abstract presentation provided above in §2.1.5 covers a single-dimensional map, $X : \Omega \to \mathbb{R}$. It is straightforward to extend this formulation to cover multidimensional versions as well. By defining

$$\text{random vector: } \mathbf{X} = (X_1, \cdots, X_n) \tag{18}$$

$$\text{multidimensional Borel set: } \mathbf{B}^n = (\mathcal{B}_1, \cdots, \mathcal{B}_n) \in \mathcal{B}^n \tag{19}$$

we can generate the multidimensional version of (14) as

$$\mathbf{X} : \Omega \to \mathbb{R}^n \quad \begin{cases} \text{Realization of the random vector:} & \omega \in \Omega \Rightarrow \mathbf{X}(\omega) = \mathbf{x} \\ \text{Borel sets pull back to measurable collections of events:} & \mathbf{X}^{-1} : \mathcal{B}^n \to \mathcal{F} \\ \text{The probability 'law of } \mathbf{X}\text{':} & \left.\begin{array}{l} \mu_{\mathbf{X}} : \mathcal{B}^n \to [0,1] \\ \mathbf{B} \in \mathcal{B}^n \end{array}\right\} \Rightarrow \mu_{\mathbf{X}}\mathbf{B} = \mathbb{P}_{\mathbf{X}}\mathbf{X}^{-1}\mathbf{B} \end{cases} \tag{20}$$

Defining the multi-dimensional half intervals, and the associated probability measures

$$\left.\begin{array}{l} \mathbf{B} = ([-\infty, x_1), \cdots, [-\infty, x_n)) \\ A = \mathbf{X}^{-1}\mathbf{B} = \{\omega : (X_1(\omega) < x_1, \cdots, X_n(\omega) < x_n)\} \end{array}\right\} \Rightarrow \mathbb{P}_{\mathbf{X}} A = \mu_{\mathbf{X}}\mathbf{B} \equiv F_{\mathbf{X}}(\mathbf{x}) \tag{21}$$

the cumulative and density functions are expressed as

$$\textbf{Probability density function: } p(\mathbf{x}) \equiv p(x_1, \cdots, x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F(x_1, \cdots, x_n) \tag{22}$$

$$\textbf{Cumulative distribution function: } F(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} p(x_1, \cdots, x_n)\, dx_1 \cdots dx_n \tag{23}$$

The random vector is a model for joint distributions over finite-dimensional spaces. Usually, if the sequence of random variables is important – as it is for a sample distribution – the joint distribution

is writte as a tuple with variables expressed as coordinates. If the sequence is not important, then the variables are expressed with different characters without enclosing parentheses: X, Y vs. (X_1, X_2). The mathematical operations for these two interpretations are identical.

For the case in which the random vector is indexed by an *uncountable continuum* of values, $0 \leq t \leq \infty$, for example, the object is termed a **random process**, a topic that is covered in a companion set of notes.

2.1.8 Independent Joint Distributions

A bivariate distribution of random variables, X and Y, is **independent** if the joint probability densities factor,

$$p_{X,Y}(x, y) = p_X(x)p_Y(y). \tag{24}$$

This definition can be extended to cover multivariate distribution of arbitrary random variables, represented as the sequence, $\mathbf{X} = (X_1, \cdots, X_n)$, with independence defined through

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} p_{X_i}(x_i). \tag{25}$$

As a special case a multivariate distribution is **independent identically distributed (IID)** provided that the coordinates of the random vector are distributed by a *common* random variable,

$$X_1, \ldots, X_n \sim X \Rightarrow p_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} p_X(x_i) \tag{26}$$

2.1.9 Joint, Conditional and Marginal Distributions

A multivariate joint distribution, independent or otherwise, can be factored into **marginal** and **conditional** distributions

$$\left.\begin{array}{ll} \text{joint distribution:} & p_{X,Y}(x, y) \\ \text{marginal distribution:} & p_Y(y) \\ \text{conditional distribution:} & p_{X|Y}(x|y) \end{array}\right\} \Rightarrow p_{X,Y}(x, y) = p_{X|Y}(x|y)\, p_Y(y). \tag{27}$$

It is frequently useful to express the conditional distribution in terms of the joint and marginal distributions,

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}. \tag{28}$$

Notice that independence in the random variables, X and Y, by the factorization in (24), requires that the conditional distribution coincide with the marginal distribution, since

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} = \frac{p_X(x)\, p_Y(y)}{p_Y(y)} = p_X(x). \tag{29}$$

Finally, since the conditional operation is symmetric in the random variables, X and Y, we can rearrange the terms,

$$p_{X|Y}(x|y)p_Y(y) = p_{X,Y}(x, y) = p_{Y|X}(y|x)p_X(x) \Rightarrow p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)}, \tag{30}$$

for which the latter formation is known as **Bayes' Theorem** and is treated in greater detail below in §11.

2.1.10 The Expectation Operator

The main tool in probability theory is the **expection operator**, $\mathbb{E}$, which carries out integration of constants, functions of the random variable, X, or functions of the probability measure, $\mathbb{P}_X$, taken over the entire probability space. In all cases the integration over the probability space is equal to the integration of the random variable mapped into the state space, and weighted by the probability density function. An arbitrary function of the random variable, for example, is expressed as

$$\mathbb{E}gX = \int_{-\infty}^{\infty} g(x)\, p_X(x)\, dx. \tag{31}$$

In fact *all* operations in this presentation can be expressed as applications of the expectation operator. Note in particular that probability calculations can be expressed in terms of the expectation operator, since

$$1_A(\omega) = \begin{cases} 1, \omega \in A \\ 0, \omega \notin A \end{cases} \quad \Rightarrow \mathbb{P}A = \mathbb{E}\, 1_A, \tag{32}$$

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

3 Moments

3.1 Univariate

A (continuous) probability distribution is completely defined by its **moments**, which are integrations of powers of the random variable,

$$\mathbb{E}X^n = \int_{-\infty}^{\infty} x^n\, p_X(x)\, dx. \tag{33}$$

A common method of approximating probability distributions is to adjust parameters to match the measured values of the lower-order moments. The first moment is known as the **mean** of the distribution, and is given by

$$\mathbb{E}X \equiv \mu_X = \int_{-\infty}^{\infty} x\, p_X(x)\, dx. \tag{34}$$

Note that the standard symbol for the mean, μ_X, matches the form typically assigned to the probability measure in the state space. It should be obvious from context which meaning is intended.

Information about the shape of the distribution is also provided by the **central moments**, which are integrations of powers of the random variable shifted by the mean,

$$\mathbb{E}(X - \mathbb{E}X)^n = \int_{-\infty}^{\infty} (x - \mu_X)^n\, p_X(x)\, dx. \tag{35}$$

In particular the second central moment, also known as the **variance**, is used as a common measure of spread of a distribution,

$$\mathbb{V}X \equiv \sigma_X^2 = \mathbb{E}(X - \mathbb{E}X)^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 \, p(x) \, dx \tag{36}$$

$$= \mathbb{E}X^2 - (\mathbb{E}X)^2 = \int_{-\infty}^{\infty} x^2 \, p(x) \, dx - \mu_X^2 \tag{37}$$

and the covariance of a joint distribution is used a common measure of coordination of variation,

$$\mathbb{C}(X, Y \equiv \sigma_{XY} = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) \, p(x, y) \, dx \, dy \tag{38}$$

$$= \mathbb{E}XY - \mathbb{E}X \, \mathbb{E}Y = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \, p(x, y) \, dx \, dy - \mu_X \mu_Y \tag{39}$$

3.2 Multivariate Moments

The moments for multivariate probability distributions can be expressed in terms of vectors and operations on vectors of univariate random variables. In particular the mean and covariance of the multivariate distribution take the form of a vector and matrix, respectively:

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \Rightarrow \begin{cases} \mathbb{E}\mathbf{X} & = \begin{pmatrix} \mathbb{E}X_1 \\ \vdots \\ \mathbb{E}X_n \end{pmatrix} \\ \mathbb{V}\mathbf{X} & = \mathbb{E}(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})^\top = \mathbb{E}\mathbf{X}\mathbf{X}^\top - \mathbb{E}\mathbf{X} \, \mathbb{E}\mathbf{X}^\top \\ & = \begin{pmatrix} \mathbb{V}X_1 & \cdots & \mathbb{C}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \mathbb{C}(X_n, X_1) & \cdots & \mathbb{V}X_n \end{pmatrix} \end{cases} \tag{40}$$

Note in particular that the covariance matrix is written as the expectation of a random variable dyad (see LAN, §3.3), and is generally full-rank, symmetric and positive definite.

3.3 Sample Mean and Variance

3.3.1 Single-Variate Points

Given a set of data points, $(x_1, \cdots, x_m)$, the sample mean and sample variance are given by

$$\text{sample mean:} \quad \bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i \tag{41}$$

$$\text{sample variance:} \quad s_x = \frac{1}{m} \sum_{i=1}^{m} (x_i - \bar{x})^2 = \frac{1}{m} \sum_{i=1}^{m} x_i^2 - \bar{x}^2. \tag{42}$$

Representing the data points as a vector, $(x_1, \cdots, x_m)^\top \equiv \mathbf{x}$, we can form the quantities,

$$\mathbf{x}^\top \mathbf{1}_m = m\bar{x}, \tag{43}$$

$$\mathbf{x}^\top \mathbf{1}_m (\mathbf{1}_m^\top \mathbf{1}_m)^{-1} \mathbf{1}_m^\top \mathbf{x} = \mathbf{x}^\top P_{\mathbf{1}_m} \mathbf{x} = m\bar{x}^2 \tag{44}$$

for which the matrix, $P_{\mathbf{1}_m}$, is a projection matrix (*cf. LAN*, §3.11) into the 1-dimensional subspace spanned by the ones vector. The sample mean and sample variance can then be expressed as,

$$\text{sample mean:} \quad \bar{x} = \frac{1}{m}\mathbf{x}^\top \mathbf{1}_m \tag{45}$$

$$\text{sample variance:} \quad s_x = \frac{1}{m}\mathbf{x}^\top \mathbf{x} - \frac{1}{m}\mathbf{x}^\top P_{\mathbf{1}_m}\mathbf{x} = \frac{1}{m}\mathbf{x}^\top (I_m - P_{\mathbf{1}_m})\mathbf{x} \tag{46}$$

$$= \frac{1}{m}[(I_m - P_{\mathbf{1}_m})\mathbf{x}]^\top [(I_m - P_{\mathbf{1}_m})\mathbf{x}]. \tag{47}$$

Note that the sample mean and variance lie in orthogonal 1- and $(m-1)$-dimensional subspaces, respectively.

Also note that the normalizing factor for the sample variance – here, the number of points, m – will be shown to yield a biased estimator for variance that maximizes entropy; an unbiased estimator is achieved by normalizing the sum by $(m-1)$.

3.3.2 Multivariate Points

For a set of multidimensional data points, $(\mathbf{x}_1, \cdots, \mathbf{x}_m)$, each taken from an n-dimensional space, we can arrange the information in matrix form so that each row contains a single data point,

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix} = \begin{pmatrix} \leftarrow & \mathbf{x}_1 & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_m & \rightarrow \end{pmatrix} \equiv \begin{pmatrix} \uparrow & & \uparrow \\ \mathbf{c}_1 & \cdots & \mathbf{c}_n \\ \downarrow & & \downarrow \end{pmatrix}. \tag{48}$$

and the last equivalence stresses that the *coordinates* are aligned in columns. Notice also that the symbol, X, is a *matrix*, not a random variable. Using the sample mean and sample variance formulas in (41) and (42), and defining the m-dimensional ones vector, $\mathbf{1}_m = \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}^\top$, we can express the *coordinate-wise* sample means and variances in terms of inner products (*cf. LAN*, §3.2),

$$\text{coordinate sample mean:} \quad \bar{x}_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij} = \frac{1}{m}\mathbf{c}_i^\top \mathbf{1}_m \tag{49}$$

$$\text{coordinate sample covariance:} \quad s_{jk} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) = \frac{1}{m}\mathbf{c}_i^\top \mathbf{c}_j - \bar{x}_i\bar{x}_j \tag{50}$$

Given the following matrix operations,

$$X^\top \mathbf{1}_m = m \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_n \end{pmatrix}, \tag{51}$$

$$X^\top \mathbf{1}_m (\mathbf{1}_m^\top \mathbf{1}_m)^{-1} \mathbf{1}_m^\top X = X^\top P_{\mathbf{1}_m} X = m \begin{pmatrix} \bar{x}_1^2 & \cdots & \bar{x}_1\bar{x}_n \\ \vdots & \ddots & \vdots \\ \bar{x}_n\bar{x}_1 & \cdots & \bar{x}_n^2 \end{pmatrix}, \tag{52}$$

for which the matrix, $P_{\mathbf{1}_m}$, is a projection matrix (*cf. LAN*, §3.11). The sample mean (vector) and sample variance (covariance matrix) can be expressed in vector and matrix form,

$$\text{sample mean vector:} \quad \bar{\mathbf{x}} = \frac{1}{m}X^\top \mathbf{1}_m \tag{53}$$

$$\text{sample covariance matrix:} \quad S_X = \frac{1}{m}X^\top X - \frac{1}{m}X^\top P_{\mathbf{1}_m} X = \frac{1}{m}X^\top (I_m - P_{\mathbf{1}_m})X \tag{54}$$

$$= \frac{1}{m}[(I_m - P_{\mathbf{1}_m})X]^\top [(I_m - P_{\mathbf{1}_m})X]. \tag{55}$$

Notice that the information in the sample mean vector and sample covariance matrix lie in the ranges of the projection operators, $P_{\mathbf{1}_m}$ and $(I_m - P_{\mathbf{1}_m})$, that are applied to $m \times n$ matrices, which are orthogonal m- and $m(n-1)$-dimensional subspaces, respectively.

4 Common Inequalities

Proofs of probabilistic claims frequently rely on well-established inequalities, especially when applied asymptotically. A few of the most common are presented in the following subsections.

4.1 Cauchy-Schwarz Inequality

In linear algebra the **Cauchy-Schwarz Inequality** is a statement about inner products and vector norms, (*cf. LAN*, §3.5, (43)), so that given two vectors, $\mathbf{x}$ and $\mathbf{y}$, we have

$$|\langle \mathbf{x}, \mathbf{y} \rangle|^2 \leq \langle \mathbf{x}, \mathbf{x} \rangle \cdot \langle \mathbf{y}, \mathbf{y} \rangle. \tag{56}$$

The probabilistic interpretation of the statement follows from the identification of an inner product with the expectation of the product of random variables,

$$\langle X, Y \rangle \equiv \mathbb{E}XY, \tag{57}$$

for which the bilinear operation in finite dimensions is extended to cover integrable functions defined over the real number line, $\mathbb{R}$. From this the inequality follows,

$$|\mathbb{E}XY|^2 \leq \mathbb{E}X^2 \cdot \mathbb{E}Y^2 \Rightarrow \mathbb{C}(X, Y) \leq \mathbb{V}X \cdot \mathbb{V}Y \tag{58}$$

4.2 Chebyshev's Inequality

Since the total weight of a probability distribution is unity, the density function for a random variable must asymptote to zero for large positive and negative values. If, in addition the mean of the distribution is well-defined, it is possible to bound the weight in the tails of the distribution, not only for the random variable, X, but for arbitrary non-negative functions of the random variable, $g(X) \geq 0$, via **Chebyshev's Inequality**,

$$\mathbb{P}\{g(X) \geq r\} \leq \frac{\mathbb{E}g(X)}{r}. \tag{59}$$

This follows directly from simple properties of integrals, since

$$\mathbb{E}g(X) = \int_D g(x) f(x)\, dx \geq \int_{\{g(x) \geq r\}} g(x) f(x)\, dx \geq r \int_{\{g(x) \geq r\}} f(x)\, dx = r\, \mathbb{P}\{g(X) \geq r\}. \tag{60}$$

4.3 Jensen's Inequality

Given a random variable, X, and a *convex* function, ϕ, for which

$$\left.\begin{array}{l} x_1 \leq x_2 \\ 0 \leq t \leq 1 \end{array}\right\} \Rightarrow \phi(tx_1 + (1-t)x_2) \leq t\phi(x_1) + (1-t)\phi(x_2), \tag{61}$$

the relative magnitudes of the results of applying the expectation and function operators is given by **Jensen's Inequality**,

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X). \tag{62}$$

The conditions for convexity given in (61) can be expressed globally as an inequality relation between a function and its tangent line at an arbitrary point, x, so that we have

$$ax + b \leq \phi(x), \tag{63}$$

for the appropriate values, a and b. If we choose the particular point, $x_0 = \mathbb{E}X$, then the inequality in (63) leads to the relations,

$$\mathbb{E}\phi(X) = \int_{\mathbb{R}} \phi(x)p(x)\,dx \geq \int_{\mathbb{R}} (ax+b)p(x)\,dx = \\ a\int_{\mathbb{R}} xp(x)\,dx + b\int_{\mathbb{R}} p(x)\,dx = ax_0 + b = \phi(x_0) = \phi(\mathbb{E}X). \tag{64}$$

5 Operators

All information in the random variable, X, is contained within the induced probability measure, $\mathbb{P}_X$. Other formulations of the random variable, provided below in §§5.1.1 - 5.1.3, contain equivalent information, however, organized as countably-infinite applications of the expectation operator. These alternative formulations form the basis of many proofs and demonstrations, in which a random variable is operated upon and the form of a new random variable is generated immediately or emerges asymptotically in the limit of infinite operations.

5.1 Exponentiated Operators

There are a number of useful operations on random variables that can be expressed as power series with factorial weights. This can be represented symbolically as an **exponentiated function**,

$$X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots = \sum_{n=1}^{\infty} \frac{X^n}{n!} \equiv e^X, \tag{65}$$

which is well-defined provided the moments exist.

5.1.1 Moment-Generating Functions

The **moment-generating function** is the expectation of a parametrized exponentiated function of the random variable, X, defined as

$$M_X(t) \equiv \mathbb{E}e^{tX} = \sum_{n=1}^{\infty} \frac{\mathbb{E}X^n}{n!} t^n, \tag{66}$$

which takes the form of a Taylor's series expanded about the origin. The function receives its name from the term-by-term evaluation,

$$\mathbb{E}X^n = \frac{d^n}{dt^n} M_X(t)\bigg|_{t=0} = M_X^{(n)}(0), \tag{67}$$

for which the n^{th} term in the series is the n^{th} moment of the distribution. Notice that the expression in (66) is the *Laplace transform* of the random variable.

The moment-generating function of a scaled random variable, cX, is identical to the moment-generating function with a scaled parameter,

$$M_{cX}(t) = \mathbb{E}e^{ctX} = M_X(ct) \tag{68}$$

while the joint moment-generating function for independent random variables, X and Y, (*cf.* §2.1.8), factors into the product of the marginal moment-generating functions,

$$M_{X,Y}(s,t) = \mathbb{E}e^{sX+tY} = \mathbb{E}e^{sX}e^{tY} = \mathbb{E}e^{sX}\mathbb{E}e^{tY} = M_X(s)M_Y(t). \tag{69}$$

The first equality in the chain in (69) is the definition of joint moment-generating function; the third equality is due to independence (the factorability of the joint probability measures is factored into the product of marginals).

It can be proven that continuous probability distributions are uniquely specified by their moments. It is therefore a common strategy to identify the probability distribution associated with a transformed random variable by recognizing the form of the transformed moment-generating function.

5.1.2 Characteristic Functions

The moment-generating function is the Laplace transform of the random variable; the **characteristic** function is the *Fourier transform*,

$$\phi_X(t) \equiv \mathbb{E}e^{itX}. \tag{70}$$

The characteristic function is a *globally* convergent series, unlike the moment-generating function.

As with the moment-generating function, the characteristic function of a scaled random variable is the characteristic function of the scaled parameter,

$$\phi_{cX}(t) = \mathbb{E}e^{ictX} = \phi_X(ct) \tag{71}$$

while for independent random variables, X and Y, the joint characteristic function is factored into the product of marginals,

$$\phi_{X,Y}(s,t) = \mathbb{E}e^{i(sX+tY)} = \mathbb{E}e^{isX}e^{itY} = \mathbb{E}e^{isX}\mathbb{E}e^{itY} = \phi_X(s)\phi_Y(t). \tag{72}$$

5.1.3 Cumulants

The logarithm of the moment-generating function, called the **cumulant** function,

$$K_X(t) = \ln M_X(t) \tag{73}$$

is another function of a random variable whose information is equivalent to the probability density function. Here, we can expand the logarithm about unity, $\ln(1+x) = t - \frac{t^2}{2} + \cdots$, to derive the first two terms of the infinite series,

$$\begin{aligned} K_X(t) &= \left(t\mathbb{E}X + \frac{t^2}{2}\mathbb{E}X^2 + \cdots\right) + \frac{1}{2}\left(t\mathbb{E}X + \cdots\right)^2 + \cdots \\ &= t\mathbb{E}X + \frac{t^2}{2}\left(\left(\mathbb{E}X\right)^2 - \mathbb{E}X^2\right) + \cdots . \end{aligned} \tag{74}$$

The second term in the cumulant expansion is the variance of the random variables.

As with moment-generating functions, the cumulant of the scaled random variable is the cumulant function operating on the scaled parameter,

$$K_{cX}(t) = K_X(ct), \tag{75}$$

while for independent random variables, X and Y, the joint cumulant function is expressed as the sum of the marginals,

$$K_{X,Y}(s,t) = K_X(s) + K_Y(t). \tag{76}$$

5.1.4 Extensions to Random Vectors

Each of the functions defined above – the moment-generating, characteristic, and cumulant functions – can be extended to cover multivariate distributions, represented as vectors, $\mathbf{X}$. The parameter is also modified, coverted to a vector with matching dimension, so that we have,

$$\left.\begin{array}{l} X \to \mathbf{X} \\ t \to \mathbf{t} \end{array}\right\} \Rightarrow \begin{cases} M_{\mathbf{X}}(\mathbf{t}) \equiv \mathbb{E}e^{\mathbf{t}^\top \mathbf{X}} \\ K_{\mathbf{X}}(\mathbf{t}) \equiv \ln M_{\mathbf{X}}(\mathbf{t}) \\ \phi_{\mathbf{X}}(\mathbf{t}) \equiv \mathbb{E}e^{i\mathbf{t}^\top \mathbf{X}} \end{cases} \tag{77}$$

Properties due to scaling and to application to independent coordinate random variables carry over as expected.

5.2 Transformations

5.2.1 General Transformation

$$Y = g(X) \tag{78}$$

$$\text{increasing function, } g : F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq Y = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) \tag{79}$$

$$\text{decreasing function, } g : F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq Y = \mathbb{P}(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)) \tag{80}$$

$$p_Y(y) = \frac{d}{dy} F_Y(y) = p_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \tag{81}$$

5.2.2 Scale-location Adjustment

$$X \sim p(x) \Rightarrow \alpha + \beta X \sim \frac{1}{\beta} p(\alpha + \beta x) \tag{82}$$

5.2.3 Sum of Random Variables

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

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) \Rightarrow p_{X+Y}(z) = \int_{-\infty}^{\infty} p_X(x) p_Y(z - x)\, dx \tag{83}$$

6 Common Functions

Many of the common distributions used as continuous probabilistic models, described in detail below in §7.3, can be expressed in closed form using functions defined through definite integrals.

6.1 The Error Function

The **error function** and **complementary error function** are defined as,

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt \tag{84}$$

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) \tag{85}$$

which are used for closed-form expression of cumulative Gaussian probabilities, as shown in §7.2.1. Notice that the argument, x, appears as the limit of integration.

6.2 The Gamma Function

The **gamma function** is a generalization of the factorial function, extending the application from positive integers to the entire real number line. The function is defined as

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt, \tag{86}$$

for which the argument, x, appears as a parameter within the definite integral. The recursive property of the function is demonstrated though integration by parts,

$$\begin{aligned}
\Gamma(x) &= \int_0^\infty t^{x-1} e^{-t}\, dt \\
&= -t^{x-1} e^{-t}\big|_0^\infty - \int_0^\infty (x-1) t^{x-2} (-e^{-t})\, dt \\
&= (x-1) \int_0^\infty t^{x-2} e^{-t}\, dt \\
&= (x-1)\Gamma(x-1).
\end{aligned} \tag{87}$$

By restricting the argument to integral values, $x = n$, we recover the factorial relation,

$$\Gamma(n) = (n-1)! \tag{88}$$

6.3 The Beta Function

The **beta function** is bivariate function defined through the definite integral,

$$\mathrm{B}(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt \tag{89}$$

for which the arguments appear as parameters in the integrand. The beta function can also be defined through gamma functions, since

$$\begin{aligned}
\Gamma(x)\Gamma(y) &= \int_0^\infty s^{x-1} e^{-s}\, ds \int_0^\infty t^{y-1} e^{-t}\, dt \\
&= \int_0^\infty \int_0^\infty s^{x-1} t^{y-1} e^{-(s+t)}\, ds\, dt && \begin{cases} s = uv \\ t = u(1-v) \end{cases} \\
&= \int_{u=0}^\infty \int_{v=0}^1 (uv)^{x-1} (u(1-v))^{y-1} e^{-u} u\, du\, dv && |J| = u \\
&= \int_0^\infty e^{-u} u^{x+y-1}\, du \int_0^1 v^{x-1} (1-v)^{y-1}\, dv \\
&= \Gamma(x+y)\, \mathrm{B}(x, y)
\end{aligned} \tag{90}$$

which can be rearrange to yield,

$$\mathrm{B}(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}. \tag{91}$$

As a special case, the beta function evaluated at the points, $x = y = \frac{1}{2}$,

$$\begin{aligned}\mathrm{B}\left(\frac{1}{2},\frac{1}{2}\right) &= \int_0^1 t^{-\frac{1}{2}}(1-t)^{-\frac{1}{2}}\,dt \\ &= 2\int_0^{\frac{\pi}{2}} \frac{\cos\theta\sin\theta}{\cos\theta\sin\theta}\,d\theta \\ &= \pi\end{aligned} \tag{92}$$

can also be used to calculate the gamma function at the point, $x = \frac{1}{2}$,

$$\mathrm{B}\left(\frac{1}{2},\frac{1}{2}\right) = \frac{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(1)} \Rightarrow \Gamma\left(\frac{1}{2}\right) = \sqrt{\mathrm{B}\left(\frac{1}{2},\frac{1}{2}\right)} = \sqrt{\pi}. \tag{93}$$

Finally, notice that the combinatorial function can be expressed as in terms of the beta function,

$$\binom{n}{k} \equiv \frac{n!}{k!\,(n-k)!} \equiv \frac{1}{n+1}\frac{1}{\mathrm{B}(k+1,n-k+1)}. \tag{94}$$

6.3.1 The Multivariate Beta Function

The relation between beta and gamma functions in (91) can be generalized for multivariate argument,

$$\mathrm{B}(\alpha_1,\ldots,\alpha_n) = \frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^n \alpha_i\right)}. \tag{95}$$

This can be alternatively expressed in terms of a beta function of an arbitrary argument, and the sum of the remainder, as in

$$\mathrm{B}(\alpha_1,\ldots,\alpha_n) = \frac{\Gamma(\alpha_j)\prod_{i\neq j}\Gamma(\alpha_i)}{\Gamma\left(\alpha_j + \sum_{i\neq j}\alpha_i\right)} = \frac{\prod_{i\neq j}\Gamma(\alpha_i)}{\Gamma(\sum_{i\neq j}\alpha_i)}\,\mathrm{B}\left(\alpha_j, \sum_{i\neq j}\alpha_i\right). \tag{96}$$

7 Common Distributions

7.1 Discrete Distributions

Discrete distributions are frequently used to model IID sampling from populations with a finite number of types. The type counts may be two or many, the populations may be infinite (with replacement) or finite (without replacement), there may be constraints on the total number or position of specific sample types, but distinct sample values are uncorrelated in all distributions treated here.

For the distributions sampled from binary populations – two types, typically 'true' or 'false' – the weights can be taken from positions within horizontal or diagonal rows in Pascal's triangle. Or similarly, all can be ultimately expressed as combinations of binary samples, each of which is a **Bernoulli trial**.

7.1.1 Sampling With Replacement

7.1.1.1 Bernoulli

The Bernoulli distribution is derived from a single sample from a binary distribution for which the positive outcome is given a value, $0 \leq p \leq 1$. This can be expressed succintly as

$$\text{Ber}(p) \equiv f(k|p) = p^k(1-p)^{1-k}, k \in \{0, 1\} \tag{97}$$

The low-order moments are given by

$$\text{mean: } p \tag{98}$$

$$\text{variance: } p(1-p) \tag{99}$$

This is the very tip of Pascal's triangle.

7.1.1.2 Binomial

The binomial distribution provides the estimate of k successes in n samples, each of which is a Bernoulli trial, so that

$$X_i \sim \text{Ber}(p) \Rightarrow Y = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, p) \tag{100}$$

The form of the density function is the product of success and failure probabilities, weighted by the total number of combinations:

$$\text{Bin}(n, p) \equiv f(k|n, p) = \binom{n}{k} p^k(1-p)^{n-k}, k \in \{0, \cdots, n\} \tag{101}$$

The low-order moments are given by

$$\text{mean: } np \tag{102}$$

$$\text{variance: } np(1-p) \tag{103}$$

These coefficients are supplied by the n^{th} horizontal row of Pascal's triangle.

7.1.1.3 Negative Binomial

The negative binomial distribution covers the case in which k successes and r failures are recorded among $n = k + r$ independent Bernoulli trials, and the last sample is failure. This combination can be expressed as the product of binomial and Bernoulli events, expressed as

$$\text{NB}(k|r, p) = \text{Bin}(k|k + r - 1, p)\,\text{Ber}(0|p) \tag{104}$$

and so the density function takes the form.

$$\text{NB}(k|r, p) \equiv f(k|r, p) = \binom{k + r - 1}{k} p^k(1-p)^r, k \in \mathbb{N}. \tag{105}$$

The low-order moments are given by

$$\text{mean: } \frac{rp}{1-p} \tag{106}$$

$$\text{variance: } \frac{rp}{(1-p)^2} \tag{107}$$

Note that the coefficients are supplied by the r^{th} diagonal column of Pascal's triangle.

7.1.1.4 Geometric

The geometric distribution is a special case of the negative binomial distribution, for which success and failure are interchanged and a single success is recorded:

$$\mathrm{Geo}(n|p) = \mathrm{NB}(n-1|1,1-p) = p(1-p)^{n-1}, n \in \mathbb{N}. \tag{108}$$

The low-order moments are given by

$$\text{mean: } \frac{1-p}{p} \tag{109}$$

$$\text{variance: } \frac{1-p}{p^2} \tag{110}$$

Here, for consistency with definitions of the binomial and negative binomial distributions the variable that enters into the density formula is n, defined as number of trials.

7.1.1.5 Poisson

The Poisson distribution is a discrete distribution with a continuous weight,

$$\mathrm{Poi}(\lambda) = f(k|\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}, k \in \mathbb{N}. \tag{111}$$

The distribution can be derived from the binomial distribution in the asymptotic limit as the parameter, $\lambda = np$, is the product of sample size and success probability, and the number of samples diverges. For any finite sample size we have

$$\begin{aligned}
\mathrm{Bin}\left(n, \frac{\lambda}{n}\right) &= f\left(k \middle| n, \frac{\lambda}{n}\right) = \binom{n}{k}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{k}\right)^{n-k} \\
&= \frac{n \cdot n-1 \cdot \cdots \cdot n-k+1}{k!}\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^k \\
&= \left[\left(\frac{n}{n}\right) \cdot \left(\frac{n-1}{n}\right) \cdot \cdots \cdot \left(\frac{n-k+1}{n}\right)\right] \cdot \left[\left(1-\frac{\lambda}{n}\right)^{-k}\right] \cdot \left[\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\right]
\end{aligned} \tag{112}$$

which yields the Poisson distribution in the limit,

$$\lim_{n\to\infty} \mathrm{Bin}\left(n, \frac{\lambda}{n}\right) = \frac{\lambda^k}{k!}e^{-\lambda} \equiv \mathrm{Poi}(\lambda). \tag{113}$$

The low-order moments are given by

$$\text{mean: } \lambda \tag{114}$$

$$\text{variance: } \lambda \tag{115}$$

7.1.1.6 Multinomial

The multinomial distribution is realized by the sum of n repeated *composite* Bernoulli trials, for which the number of possible outcome categories is extended from two to k. Although the outcome of each composite Bernoulli trial is independent of the others, the outcome probabilities taken over all categories are linked by the requirement that one, and only one, is successful on any given trial:

$$\mathbf{p} = (p_1, \cdots, p_k)^\top, \quad \sum_{i=1}^{k} p_i = 1. \tag{116}$$

Given this underlying vector of outcome probabilities, the multinomial density function is expressed as

$$\text{Mul}(n, p_1, \cdots, p_k) \equiv f(x_1, \cdots, x_k | n, p_1, \cdots, p_k) = \frac{n}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i} = \frac{\Gamma(1 + \sum_{i=1}^{k} x_i)}{\prod_{i=1}^{k} \Gamma(1 + x_i)} \prod_{i=1}^{k} p_i^{x_i} \tag{117}$$

for which the low-order moments of the coordinate random variables are given by

$$\mathbf{X}_n \equiv (X_1, \cdots, X_k)^\top \sim \text{Mul}(n, p_1, \cdots, p_k) \Rightarrow \begin{cases} \mathbb{E}X_i = np_i \\ \mathbb{V}X_i = np_i(1 - p_i) \\ \mathbb{C}(X_i, X_j) = -np_i p_j, i \neq j \end{cases} \tag{118}$$

It is possible to express the binomial distribution, (*cf.* 7.1.1.2), as a special case of the multinomial distribution for which there are only two possible outcomes on any given trial, $k = 2$.

7.1.2 Sampling Without Replacement

7.1.2.1 Hypergeometric

The **hypergeometric distribution** is closely related to the binomial distribution, with samples taken from a finite population without replacement. Here, the total population and number of possible successes are given by N and K, respectively, while the number of samples and sampled successes are represented as n and k. The distribution is defined as

$$\text{Hyp}(n, N, K) \equiv f(k | n, N, K) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{119}$$

The low-order moments are given by

$$\text{mean: } n\frac{K}{N} \tag{120}$$

$$\text{variance: } n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1} \tag{121}$$

7.1.2.2 Multivariate Hypergeometric

The **multivariate hypergeometric distribution** is the multivariate version of the multinomial distribution. For the case of m categories the distribution is given by

$$\left.\begin{array}{l} \mathbf{k} = (k_1, \ldots, k_m)^\top \\ \mathbf{K} = (K_1, \ldots, K_m)^\top \end{array}\right\} \Rightarrow \text{MHG}(\mathbf{k}|\mathbf{K}) = \frac{\binom{K_1}{k_1} \cdots \binom{K_m}{k_m}}{\binom{\sum_{i=1}^{m} K_i}{\sum_{i=1}^{m} k_i}} \tag{122}$$

7.2 Continuous Distributions

We collect a few of the most common continuous distribution in the following sections, showing

- probability density and cumulative distribution functions
- moment-generating and/or characteristic functions
- mean and variance
- sums of random variables
- mulitvariate versions.

7.2.1 Gaussian Distributions

The Gaussian distribution, also called the **normal distribution**, is perhaps the most important distribution of all, governing asymptotic distributions of sample means through the central limit theorem, as discussed in §9.2.2. Since many phenomena are composed of small, additive processes, the Gaussian distribution serves well as a general model. Many other distributions, such as the chi-square distribution, T-distribution and F-distribution, are derived from transformed Gaussian random variables, and lead ultimately to many common statistical tests.

7.2.1.1 Univariate Gaussian

The univariate Gaussian distribution is defined by two parameters, μ and σ^2, which specify the mean and variance of the distribution, respectively. The probability density and cumulative distribution functions for the univariate Gaussian distribution are given by

$$\text{probability density:} \quad \mathrm{N}(\mu,\sigma^2) \equiv p_{\mathrm{N}}(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{123}$$

$$\text{cumulative distribution:} \quad F_{\mathrm{N}}(x|\mu,\sigma^2) = \int_{-\infty}^{x} p_{\mathrm{N}}(x|\mu,\sigma^2)\,dx = \frac{1}{2}\left(1+\operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right) \tag{124}$$

The moments of the Gaussian distribution can be calculated by way of a helper function,

$$g(\alpha) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\alpha\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \frac{1}{\sqrt{\alpha}}, \tag{125}$$

and the central moments are derived by direct calcuation,

$$\mathbb{E}(X-\mu)^{2n-1} = 0 \tag{126}$$

$$\mathbb{E}(X-\mu)^{2n} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^{2n} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = (-2\sigma^2)^n \left.\frac{d^n}{d\alpha^n} g(\alpha)\right|_{\alpha=1} = (2n-1)!!\,\sigma^{2n} \tag{127}$$

Notice that, by symmetry all odd central moments vanish, and even central moments are expressed in powers of the parameter, σ^2. All information in the Gaussian distribution is derived from the parameters, μ and σ^2, which are the mean and variance of the distribution, established by assigning $n = 1$ and 2.

$$\begin{aligned} M_{\mathrm{N}(\mu,\sigma^2)}(t) \equiv \mathbb{E}e^{tX} &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2-tx}{2\sigma^2}}\,dx \\ &= e^{t\mu+\frac{1}{2}t^2\sigma^2} \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-(\mu+t\sigma^2))^2}{2\sigma^2}}\,dx \\ &= e^{t\mu+\frac{1}{2}t^2\sigma^2} \end{aligned} \tag{128}$$

$$\phi_{\mathrm{N}(\mu,\sigma^2)}(t) \equiv \mathbb{E}e^{itX} = e^{it\mu-\frac{1}{2}t^2\sigma^2} \tag{129}$$

Finally, the sums of independent Gaussian random variables is also a Gaussian random variable,

$$\left.\begin{aligned} X \sim \mathrm{N}(\mu,\sigma^2) \Rightarrow M_X = e^{t\mu+\frac{1}{2}t^2\sigma^2} \\ Y \sim \mathrm{N}(\nu,\tau^2) \Rightarrow M_Y = e^{t\nu+\frac{1}{2}t^2\tau^2} \end{aligned}\right\} \Rightarrow M_{X+Y} = e^{t(\mu+\nu)+\frac{1}{2}t^2(\sigma^2+\tau^2)} \Rightarrow X+Y \sim \mathrm{N}(\mu+\nu,\sigma^2+\tau^2). \tag{130}$$

7.2.1.2 Standard Normal

The **standard normal** distribution is a special case of the Gaussian distribution for which the mean and variance are given by $\mu = 0$ and $\sigma^2 = 1$, respectively, so that the standard normal random variable, Z, is represented as

$$Z \sim \mathrm{N}(0, 1). \tag{131}$$

All other derived quantities and functions are calculated similarly.

7.2.2 Gamma-Derived Distributions

The form of the probability density functions in this section are derived from the gamma function, shown above in §6.2. The gamma distribution is a common choice for models that restrict the domain to the positive real line.

7.2.2.1 Gamma

The probability density function for the gamma distribution is governed by two parameters, α and β, and takes the form,

$$\text{probability density: } \Gamma(\alpha, \beta) \equiv p_\Gamma(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},\ x \geq 0. \tag{132}$$

The specific form is derived from the gamma function, whose expression is a definite integral over the positive real numbers given in (86), for the parameter, α,

$$\begin{aligned} \Gamma(\alpha) &= \int_0^\infty t^{\alpha-1} e^{-t}\, dt \\ &= \int_0^\infty (\beta x)^{\alpha-1} e^{-\beta x} \beta\, dx \qquad t = \beta x \\ &= \int_0^\infty \beta^\alpha x^{\alpha-1} e^{-\beta x}\, dx \end{aligned} \tag{133}$$

and for which the parameter, β, is introduced as a scaling variable for the integration. The integrand of the density function in (132) matches the integrand of the normalizing constant, and so the total weight is unity.

It is a straightforward exercise in integration to calculate the moments of the gamma distribution, since one part of the integrand is a power of the independent variable. This leads to a rearrangement of the terms inside the integral, which can be expressed as gamma distribution governed by *a different* parameter set that must also possess unit weight:

$$\begin{aligned} \mathbb{E}X^n &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^k x^{\alpha-1} e^{-\beta x}\, dx \\ &= \frac{\beta^\alpha}{\beta^{\alpha+k}} \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \int_0^\infty \frac{\beta^{\alpha+k}}{\Gamma(\alpha+k)} x^{\alpha+k-1} e^{-\beta x}\, dx \\ &= \frac{1}{\beta^k} \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \int_0^\infty p_\Gamma(x|\alpha+k, \beta)\, dx \\ &= \frac{1}{\beta^k} \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \end{aligned} \tag{134}$$

The mean and variance are therefore given by

$$\mathbb{E}X = \frac{\alpha}{\beta}; \tag{135}$$

$$\mathbb{V}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{(\alpha+1)\alpha}{\beta^2} - \frac{\alpha^2}{\beta^2} = \frac{\alpha}{\beta^2}. \tag{136}$$

The moment-generating function is calculated through direct integration, since the other factor in the integrand is an exponential. We can therefore use the same technique of rearranging terms to form a different gamma distribution, this one for a transformation of the β-parameter,

$$\begin{aligned} M_{\Gamma(\alpha,\beta)} \equiv \mathbb{E}e^{tX} &= \int_0^\infty e^{tx} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta-t)x}\, dx \\ &= \frac{\beta^\alpha}{(\beta-t)^\alpha} \int_0^\infty \frac{(\beta-t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta-t)x}\, dx \\ &= \left(\frac{\beta}{\beta-t}\right)^\alpha \int_0^\infty f_\Gamma(x|\alpha,\beta-t)\, dx \\ &= \left(\frac{\beta}{\beta-t}\right)^\alpha. \end{aligned} \tag{137}$$

We can immediately apply the information in the moment-generating function to show that the sum of independent gamma distributions, all of which share the value of the β-parameter, is again a gamma distribution whose α-parameter is the sum of the individual α values:

$$X_i \sim \Gamma(\alpha_i, \beta) \Rightarrow \sum_{i=1}^n X_i \sim \Gamma\left(\sum_{i=1}^n \alpha_i, \beta\right). \tag{138}$$

This follows from the uniqueness of moment-generating function,

$$M_{\sum \Gamma(\alpha_i,\beta)} = \prod \left(\frac{\beta}{\beta-t}\right)^{\alpha_i} = \left(\frac{\beta}{\beta-t}\right)^{\sum \alpha_i} = M_{\Gamma(\sum \alpha_i,\beta)}. \tag{139}$$

7.2.2.2 Chi-square

Given a standard normal distribution, X, (*cf.* §7.2.1.2), the *square* of the random variable, X^2, is distributed as **chi-square**:

$$X \sim \mathrm{N}(0,1) \Rightarrow X^2 \sim \chi^2 = \Gamma\left(\frac{1}{2}, \frac{1}{2}\right), \tag{140}$$

which is a special case of the gamma distribution. This can be calculated directly from the cumulative distribution function of the standard normal, since

$$\mathbb{P}\{X^2 \le x\} = \mathbb{P}\{-\sqrt{x} \le X \le \sqrt{x}\} = \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt, \tag{141}$$

and the probability density function is calculated from the derivative,

$$\begin{aligned} \chi^2 \equiv p_{\chi^2}(x) &= \frac{d}{dx} \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt \\ &= \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x}{2}}\right)\left(\frac{1}{2} x^{-\frac{1}{2}}\right) - \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x}{2}}\right)\left(-\frac{1}{2} x^{-\frac{1}{2}}\right) \\ &= \frac{1}{\sqrt{2\pi}} x^{\frac{1}{2}-1} e^{-\frac{x}{2}} \\ &= \Gamma\left(\frac{1}{2}, \frac{1}{2}\right). \end{aligned} \tag{142}$$

Since sums of independent gamma random variables are also gamma distributed, as in (138), the sum of$n$ chi-square random variables is also gamma distributed, and is called the **chi-square distribution with n degrees of freedom**:

$$X_i \sim \mathrm{N}(0,1) \Rightarrow \sum_{i=1}^{n} X_i^2 \sim \chi_n^2 = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right). \tag{143}$$

The mean and variance of the chi-square distribution with n degrees of freedom follows directly from (135) and(136),

$$X \sim \chi_n^2 \Rightarrow \begin{cases} \mathbb{E}X = n; \\ \mathbb{V}X = \frac{n}{2}. \end{cases} \tag{144}$$

Finally, the **Mahalanobis distance**, which is the argument in the exponential of the multivariate Gaussian, is chi-square distributed, since given the eigenstructure of the covariance matrix,

$$\mathbf{\Sigma} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top = \mathbf{\Gamma}\mathbf{\Gamma}^\top \Rightarrow \mathbf{Z} = \mathbf{\Gamma}^{-1}(\mathbf{X} - \boldsymbol{\mu}), \tag{145}$$

we have

$$\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{\Sigma}) \Rightarrow (\mathbf{X} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) = \mathbf{Z}^\top \mathbf{Z} \sim \chi_n^2. \tag{146}$$

This is the multivariate extension to the chi-square derivation shown above in (142).

7.2.2.3 Inverse Gamma

Given a random variable, X, that is gamma distributed, the inverse random variable is inverse gamma distributed,

$$X \sim \Gamma(\alpha, \beta) \Rightarrow \frac{1}{X} \sim \mathrm{I}\Gamma(\alpha, \beta) \equiv p_{\mathrm{I}\Gamma}(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{x}\right)^{\alpha+1} e^{-\frac{\beta}{x}},\ x > 0. \tag{147}$$

The derivation of the form of the probability density function is a straightforward application of the formula for general transformation of random variables in (81):

$$\begin{aligned} Y = \frac{1}{X} \Rightarrow p_Y(y) &= p_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{y}\right)^{\alpha-1} e^{-\frac{\beta}{y}} \frac{1}{y^2} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{y}\right)^{\alpha+1} e^{-\frac{\beta}{y}}. \end{aligned} \tag{148}$$

7.2.3 Beta-Derived Distributions

The form of the probability density functions in this section are derived from the beta function, shown above in §6.3. The beta distribution is a common choice for models that restrict the domain to the unit interval.

7.2.3.1 Beta

The probability density function for the gamma distribution is governed by two parameters, α and β, and takes the form,

$$\text{probability density: } \mathrm{B}(\alpha,\beta) \equiv p_{\mathrm{B}}(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 \le x \le 1 \tag{149}$$

It is a straightforward exercise in integration to calculate the moments of the beta distribution, since one part of the integrand is a power of the independent variable. This leads to a rearrangement of the terms inside the integral, which can be expressed as a beta distribution governed by *a different* parameter set that must also possess unit weight:

$$\begin{aligned}
\mathbb{E}X^k &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^k x^{\alpha-1}(1-x)^{\beta-1}\,dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+k)\Gamma(\beta)}{\Gamma(\alpha+\beta+k)} \int_0^1 \frac{\Gamma(\alpha+\beta+k)}{\Gamma(\alpha+k)\Gamma(\beta)} x^{\alpha+k-1}(1-x)^{\beta-1}\,dx \\
&= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+k)} \int_0^1 p_{\mathrm{B}}(x|\alpha+k,\beta)\,dx \\
&= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+k)}.
\end{aligned} \tag{150}$$

The mean and variance of the beta distribution are therefore given by

$$\mathbb{E}X = \frac{\alpha}{\alpha+\beta}; \tag{151}$$

$$\mathbb{V}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{152}$$

7.2.3.2 Dirichlet

The **Dirichlet distribution** is a multivariate extension to the beta distribution, governed by an arbitrary number of parameters, $\alpha_1, \cdots, \alpha_n$, and takes the form

$$\text{probability density: } \mathrm{Dir}(\alpha_1,\cdots,\alpha_n) \equiv p_{\mathrm{D}}(x_1,\cdots,x_n|\alpha_1,\cdots,\alpha_n) = \frac{\prod_{i=1}^n x_i^{\alpha_i-1}}{\mathrm{B}(\alpha_1,\cdots,\alpha_n)}, \begin{cases} 0 \le x_i \le 1 \\ \sum_{i=1}^n x_i = 1 \end{cases} \tag{153}$$

Here, the normalizing constant is provided by the multivariate beta function, defined above in (95).

Notice that the marginal distributions, X_i – formed by isolating a single parameter, α_i, while lumping the remainder together, $\beta = \sum_{j\neq i} \alpha_j$ – yield beta distributions,

$$p_D(x_i|\alpha_1,\cdots,\alpha_n) = p_{\mathrm{B}}\left(x_i \middle| \alpha_i, \sum_{j\neq i} \alpha_j\right) = \frac{x_i^{\alpha_i-1}(1-x_i)^{\sum_{j\neq i}\alpha_j - 1}}{\mathrm{B}(\alpha_i, \sum_{j\neq i}\alpha_j)}. \tag{154}$$

The mean and variance are straightforward generalizations of the values for the beta distribution as well,

$$\mathbb{E}X_i = \frac{\alpha_i}{\sum_{j=1}^n \alpha_j}; \tag{155}$$

$$\mathbb{V}X_i = \frac{\alpha_i \sum_{j\neq i}\alpha_j}{\left(\sum_{j=1}^n \alpha_j\right)^2}. \tag{156}$$

7.2.4 Distributions of Ratios of Standard Normal Random Variables

This section treats distributions that arise as the ratios of standard normal variables, and serve as the basis of common test of statistical properties of the underlying distribution.

- F-distribution: ratio of average sums of squares of standard normal random variables;
- T-distribution: ratio of a single standard normal random variable to the square root of theaverage sum of squares of standard normal random variables;
- Cauchy distribution: ratio of two standard normal random variables.

Furthermore, all standard normal random variables that participate in the sums and ratios are required to be independent. And since sums of squares of standard normal random variables are themselves chi-square distributed, the F- and T-distributions can be rephrased in these terms.

7.2.4.1 F-Distribution

The **F-distribution** is defined as the ratio of average sums of squares of independent standard normal random variables,

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

$$F(k,m) \equiv p_F(x|k,m) = \frac{\Gamma(\frac{k+m}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})} \left(\frac{k}{m}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1} \left(1 + \frac{k}{m}x\right)^{-\frac{k+m}{2}} \quad (157)$$

Notice that we can define both the numerator and denominator in terms of chi-square distributed variables, which leads to a direct integral representation of the cumulative distribution function,

$$\left.\begin{matrix} U = \sum_{i=1}^{k} U_i^2 \sim \chi_k^2 \\ V = \sum_{j=1}^{m} V_j^2 \sim \chi_m^2 \end{matrix}\right\} \Rightarrow \mathbb{P}\left\{\frac{\frac{U}{k}}{\frac{V}{m}} \leq x\right\} = \mathbb{P}\{U \leq \frac{k}{m} x V\}$$

$$= \iint_{\{U \leq \frac{k}{m} x V\}} p_{\chi_k^2}(u) p_{\chi_m^2}(v)\, du\, dv = \int_0^\infty \int_0^{\frac{k}{m} x v} p_{\chi_k^2}(u) p_{\chi_m^2}(v)\, du\, dv \quad (158)$$

The probability density function is the derivative of the cumulative function, so that

$$\begin{aligned}
F(k,m) \equiv p_F(x|k,m) &= \frac{d}{dx}\int_0^\infty \int_0^{\frac{k}{m}xv} p_{\chi^2_k}(u)p_{\chi^2_m}(v)\,du\,dv = \int_0^\infty \frac{d}{dx}\left(\int_0^{\frac{k}{m}xv} p_{\chi^2_k}(u)\,du\right) p_{\chi^2_m}(v)\,dv \\
&= \frac{k}{m}\int_0^\infty p_{\chi^2_k}\left(\frac{k}{m}xv\right) p_{\chi^2_m}(v)\,v\,dv \\
&= \frac{k}{m}\int_0^\infty \left(\frac{\frac{1}{2}^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)}\left(\frac{k}{m}xv\right)^{\frac{k}{2}-1} e^{-\frac{k}{m}\frac{xv}{2}}\right)\left(\frac{\frac{1}{2}^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)}v^{\frac{m}{2}-1}e^{-\frac{v}{2}}\right) v\,dv \\
&= \frac{k^{\frac{k}{2}}}{m}\int_0^\infty \frac{\left(\frac{1}{2}\right)^{\frac{k+m}{2}} x^{\frac{k}{2}-1}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} v^{\frac{k+m}{2}-1} e^{-\frac{1}{2}v\left(\frac{k}{m}t+1\right)}\,dv \\
&= \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)}\left(\frac{k}{m}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1}\left(\frac{1}{\frac{k}{m}x+1}\right)^{\frac{k+m}{2}} \int_0^\infty \frac{\left(\frac{t+1}{2}\right)^{\frac{k+m}{2}}}{\Gamma\left(\frac{k+m}{2}\right)} v^{\frac{k+m}{2}-1} e^{-v\frac{t+1}{2}}\,dv \\
&= \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)}\left(\frac{k}{m}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1}\left(1+\frac{k}{m}x\right)^{-\frac{k+m}{2}} \int_0^\infty p_\Gamma\left(v\left|\frac{k+m}{2},\frac{t+1}{2}\right.\right) dv \\
&= \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)}\left(\frac{k}{m}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1}\left(1+\frac{k}{m}x\right)^{-\frac{k+m}{2}}
\end{aligned} \tag{159}$$

7.2.4.2 T-Distribution

The **T-distribution** is defined as the ratio of a single standard normal random variable to the square root of an average sum of squares of standard normal random variables, all of which are independent,

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

$$T(m) \equiv p_T(x|m) = \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{m}{2})}\frac{1}{\sqrt{m}}\left(1+\frac{x^2}{m}\right)^{-\frac{m+1}{2}} \tag{160}$$

It is possible to define the T-distribution in terms of square roots of chi-square random variables, and express the cumulative distribution function in terms of the F-distribution density function,

$$\left.\begin{matrix} U = U_i^2 \sim \chi^2 \\ V = \sum_{j=1}^m V_j^2 \sim \chi^2_m \end{matrix}\right\} \;\Rightarrow\; \mathbb{P}\left\{-x \le \frac{\sqrt{U}}{\sqrt{\frac{1}{m}V}} \le x\right\} \;=\; \mathbb{P}\left\{\frac{U}{\frac{1}{m}V} \le x^2\right\} \;=\; \int_0^{x^2} p_F(v|1,m)\,dv \tag{161}$$

The calculation of the density function of the T-distribution is then given by

$$\begin{aligned}
T(m) \equiv p_T(x|m) &= \frac{1}{2}\frac{d}{dx}\int_0^{x^2} p_F(v|1,m)\,dv = x\,p_F\left(x^2|1,m\right) \\
&= \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{m}{2}\right)}\frac{1}{\sqrt{m}}\left(1+\frac{x^2}{m}\right)^{-\frac{m+1}{2}}.
\end{aligned} \tag{162}$$

Note that the numerator and denominator in the expression in (160) are expressed as the ratio of standard normal to chi-square distributions, all generated independently by a standard normal source. T-distributions are most commonly applied to ratios of sample statistics – sample mean and sample (unbiased) variance – calculated from common vector of Gaussian data points. The *independence* of the two moments is guaranteed by Fisher's Theorem (*cf.* §7.3.1.5), which forms the basis of a common statistical interval test for hypothesized means (*cf. SN*, §5.2.2.1).

7.2.4.3 Cauchy

The **Cauchy distribution** is defined as the ratio of independent standard normal random variables,

$$\left.\begin{matrix} U \\ V \end{matrix}\right\} \sim Z = \mathrm{N}(0,1) \Rightarrow \frac{U}{V} \sim \mathrm{Cau}(0,1) \equiv p_C(x) = \frac{1}{\pi}\frac{1}{x^2+1}. \tag{163}$$

The cumulative distribution can be derived by direct integration of standard normal distributions over the appropriate domain,

$$\begin{aligned} \mathbb{P}\left\{\frac{U}{V} \leq x\right\} = \mathbb{P}\left\{U \leq xV\right\} &= \iint_{\{u \leq xv\}} p_{\mathrm{N}}(u|0,1) p_{\mathrm{N}}(v|0,1)\, du\, dv \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{xv} p_{\mathrm{N}}(u|0,1) p_{\mathrm{N}}(v|0,1)\, du\, dv \end{aligned} \tag{164}$$

The probability density function is then the derivative,

$$\begin{aligned} \mathrm{Cau}(0,1) \equiv f_{\mathrm{C}}(x) &= \frac{d}{dx}\mathbb{P}\left\{\frac{U}{V} \leq x\right\} = \frac{d}{dx}\int_0^{\infty}\int_0^{xv} f_{\mathrm{N}}(u|0,1) f_{\mathrm{N}}(v|0,1)\, du\, dv \\ &= \int_{-\infty}^{\infty}\left(\frac{d}{dx}\int_{-\infty}^{xv} f_{\mathrm{N}}(u|0,1)\, du\right) f_{\mathrm{N}}(v|0,1)\, dv = \int_{-\infty}^{\infty} f_{\mathrm{N}}(xv|0,1) f_{\mathrm{N}}(v|0,1)\, dv \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2v^2}{2}} \frac{1}{2\pi} e^{-\frac{v^2}{2}} v\, dv = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\frac{v^2(x^2+1)}{2}} v\, dv \\ &= \frac{1}{2\pi}\frac{1}{x^2+1}\int_{-\infty}^{\infty} e^{-t}\, dt = \frac{1}{\pi}\frac{1}{x^2+1}\int_0^{\infty} e^{-t}\, dt \\ &= \frac{1}{\pi}\frac{1}{x^2+1}. \end{aligned} \tag{165}$$

The Cauchy distribution has no moments. The mean, in particular, although symmetric, is ill-defined: since the half-integral diverges,

$$\frac{1}{\pi}\int_0^{\infty} \frac{x}{x^2+1}\, dx = \infty, \tag{166}$$

the full integral does not converge to any single value. Symmetry about the origin does, however, require the median be zero.

Finally, the Cauchy distribution is a special case of the T-distribution,

$$\mathrm{Cau}(0,1) \equiv T(1) \tag{167}$$

7.2.5 Other Common Distributions

7.2.5.1 Uniform

The probability density function for the **uniform distribution** is a constant unit value over the unit interval,

$$\mathrm{Uni}(0,1) \equiv p_{\mathrm{U}}(x) = 1, \;\; 0 \leq x \leq 1. \tag{168}$$

7.2.5.2 Exponential

The probability density function for the **exponential distribution** is a weighted exponential over the positive real number line,

$$\mathrm{Exp}(\lambda) \equiv p_{\mathrm{E}}(x|\lambda) = \lambda e^{-\lambda x}, \qquad x \geq 0. \tag{169}$$

The exponential distribution is *memoryless*, so that cumulative probabilities in future times are independent of values in past times,

$$\mathbb{P}\{x > s + t | x > s\} = \mathbb{P}\{x > t\}. \tag{170}$$

In particular, the **hazard function** is constant:

$$h(t) = \frac{p(t)}{1 - \int_0^t p(x)\,dx} = \lambda. \tag{171}$$

7.2.5.3 Pareto

The Pareto distribution is another power-law distribution in which the probability vanishes below a threshold value, α_m, and falls exponentially with parameter, α, at values above. The probability density function is given by

$$\mathrm{Par}(\alpha, x_m) \equiv p_{\mathrm{P}}(x|\alpha, x_m) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, & x \geq x_m \\ 0, & x < x_m \end{cases} \tag{172}$$

A key property of the Pareto distribution is that the distribution conditioned on exceeding a given value is itself Pareto,

$$\mathrm{Par}(\alpha, x_m | x_1 > x_m) = \mathrm{Par}(\alpha, x_1). \tag{173}$$

The lower-order moments hold a complicated relationship with the exponential parameter,

$$\text{mean: } \begin{cases} \alpha \leq 1 & \infty \\ \alpha > 1 & \frac{\alpha x_m}{\alpha - 1} \end{cases} \tag{174}$$

$$\text{variance: } \begin{cases} \alpha \leq 1 & \text{does not exist} \\ 1 < \alpha \leq 2 & \infty \\ \alpha > 2 & \left(\frac{\alpha_m}{\alpha-1}\right)^2 \frac{\alpha}{\alpha-2} \end{cases} \tag{175}$$

The hazard rate of the Pareto distribution, taken after the burn-in period, falls with the dependent parameter,

$$h(t) = \frac{\alpha}{t}. \tag{176}$$

7.2.5.4 Weibull

The Weibull distribution arises from the study of survival analysis, or hazard analysis, in which the time to failure is proportional to a power of time. The density function takes the form,

$$\mathrm{Wei}(k, \lambda) \equiv p_{\mathrm{Wei}}(x|k, \lambda) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{177}$$

The lower-order moments are expressed in terms of the gamma function,

$$\text{mean: } \lambda\,\Gamma\left(1 + \frac{1}{k}\right) \tag{178}$$

$$\text{variance: } \lambda^2 \left[\Gamma\left(1 + \frac{2}{k}\right) - \left(\Gamma\left(1 + \frac{1}{k}\right)\right)^2\right] \tag{179}$$

7.3 Continuous Distribution (Multivariate)

7.3.1 Gaussian Derived Distributions

7.3.1.1 Multivariate Gaussian

The **multivariate Gaussian** distribution is derived from the random vector, $\mathbf{X} = (X_1, \cdots, X_n)^\top$, for which each coordinate, X_i, is itself a Gaussian random variable, defined by individual means and variances, and each pair of coordinates, X_i, X_j, is linked by the covariance of the random variables. This information is organized as a mean vector, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$,

$$\mathbf{X} = \begin{pmatrix} X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2) \\ \vdots \\ X_n \sim \mathrm{N}(\mu_n, \sigma_n^2) \end{pmatrix} \Rightarrow \begin{cases} \boldsymbol{\mu} = \begin{pmatrix} \mathbb{E}X_1 \\ \vdots \\ \mathbb{E}X_n \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \\ \\ \boldsymbol{\Sigma} = \begin{pmatrix} \mathbb{V}X_1 & \cdots & \mathbb{C}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \mathbb{C}(X_1, X_n) & \cdots & \mathbb{V}X_n \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{x_1 x_n} \\ \vdots & \ddots & \vdots \\ \sigma_{x_1 x_n} & \cdots & \sigma_n^2 \end{pmatrix} \end{cases} \tag{180}$$

Note that the covariance matrix is by construction symmetric, and is further constrained to be positive definite (*cf. LAN*, §3.8).

Given the mean vector and covariance matrix defined in (180), the multivariate Gaussian distribution takes the form,

$$\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{181}$$

Note here that the argument of the exponential in the multivariate Gaussian is a quadratic form, for which constant surfaces form nested ellipsoids in the appropriately dimensioned space.

The **standard normal** distribution is a special case of the general distribution in which the random vector is 'standardized', with coordinate random variables defied by zero mean vector and indentity covariance matrix,

$$\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\mathbf{z}^\top \mathbf{z}} \tag{182}$$

7.3.1.2 Linear Transformations of Multivariate Gaussians

It is possible to generate every multivariate Gaussian distribution from a *linear transformation* of a standard normal distribution,

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{Z}, \tag{183}$$

frow which we derive

$$\left.\begin{aligned} \mathbb{E}\mathbf{X} &\equiv \mathbb{E}(\boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{Z}) = \boldsymbol{\mu} \\ \mathbb{V}\mathbf{X} &\equiv \mathbb{E}\left(\mathbf{X} - \mathbb{E}\mathbf{X}\right)\left(\mathbf{X} - \mathbb{E}\mathbf{X}\right)^\top = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top \end{aligned}\right\} \Rightarrow \mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top) \tag{184}$$

Similarly we can recover the standard normal by inverting the relationship in (183),

$$\mathbf{Z} = \boldsymbol{\Gamma}^{-1}(\mathbf{X} - \boldsymbol{\mu}), \tag{185}$$

so that

$$\left.\begin{aligned}\mathbb{E}\mathbf{Z} &\equiv \mathbb{E}\boldsymbol{\Gamma}(\mathbf{X}-\boldsymbol{\mu}) = \mathbf{0} \\ \mathbb{V}\mathbf{Z} &\equiv \mathbb{E}\left(\mathbf{Z}-\mathbb{E}\mathbf{Z}\right)\left(\mathbf{Z}-\mathbb{E}\mathbf{Z}\right)^{\top} = \mathbb{E}\mathbf{Z}\mathbf{Z}^{\top} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-\top} = \mathbf{I}\end{aligned}\right\} \Rightarrow \mathbf{Z} \sim \mathrm{N}(\mathbf{0},\mathbf{I}) \tag{186}$$

Notice in particular that *orthogonal* linear transformations convert multivariate standard normal random variables into other standard normal random variables,

$$\mathbf{X} = Q\mathbf{Z} \Rightarrow \mathbb{V}\mathbf{X} = QIQ^{\top} = I. \tag{187}$$

Finally, linear transformations map Gaussian random variables into other Gaussian random variables:

$$\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu},\Sigma) \Rightarrow A\mathbf{X} \sim \mathrm{N}(A\boldsymbol{\mu}, A\Sigma A^{\top}). \tag{188}$$

For transformation by symmetric matrices, $A = A^{\top}$, the mean vector and covariance matrix are simply updated under change of basis (*cf. LAN*, §3.9.3).

7.3.1.3 Eigenstructure of Multivariate Gaussians

We can gain greater insight into the geometric properties of multivariate Gaussian distribution by examining the eigenstructure of the covariance matrix. By the Spectral Theorem (*cf. LAN*, §4.2) the covariance matrix can be decomposed into a product of orthonormal and diagonal matrices,

$$\boldsymbol{\Sigma} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\top} \Rightarrow \boldsymbol{\Sigma}^{-1} = \mathbf{Q}\mathbf{D}^{-1}\mathbf{Q}^{\top}. \tag{189}$$

The covariance matrix is invertible provided it is of full rank, which requires that the entries in the diagonal matrix, D, be nonzero,

$$\boldsymbol{\Sigma} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\top} \Rightarrow \mathbf{D} = \mathrm{diag}(\zeta_1^2,\ldots,\zeta_n^2) \tag{190}$$

$$\boldsymbol{\Sigma}^{-1} = \mathbf{Q}\mathbf{D}^{-1}\mathbf{Q}^{\top} \Rightarrow \mathbf{D}^{-1} = \mathrm{diag}\left(\frac{1}{\zeta_1^2},\ldots,\frac{1}{\zeta_n^2}\right). \tag{191}$$

Notice that the positive-definite property has been incorporated into the specification of the diagonal matrix – the positive entries are squared values. We can recover the linear transformation in (183) by further decomposing the diagonal matrix into a square root:

$$\boldsymbol{\Sigma} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\top} = \mathbf{Q}\mathbf{D}^{\frac{1}{2}}\mathbf{D}^{\frac{\top}{2}}\mathbf{Q}^{\top} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\top} \Rightarrow \boldsymbol{\Gamma} = \mathbf{Q}\mathbf{D}^{\frac{1}{2}}, \qquad \mathbf{D}^{\frac{1}{2}} = \mathrm{diag}(|\zeta_1|,\ldots,|\zeta_n|). \tag{192}$$

If we now introduce a change of basis (*cf. LAN*, §3.9.3) to align with the orthonormal column vectors of Q, then a vector in the new coordinate system, $\mathbf{y}$, adjusted to remove the mean, takes the form

$$\mathbf{y} = \mathbf{Q}^{\top}(\mathbf{x}-\boldsymbol{\mu}), \tag{193}$$

and the quadratic form that makes up the argument of the exponential is expressed in coordinates that lie along the principal axes of the ellipsoids, so that the transformed covariance matrix decouples,

$$\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \frac{1}{2}\mathbf{y}^{\top}\mathbf{D}^{-1}\mathbf{y} = \sum_{i=1}^{n}\frac{y_i^2}{2\zeta_i^2} \tag{194}$$

Introducing the resulting change of basis into the random variable shows that the probability density function completely decouples as well,

$$\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \det\boldsymbol{\Sigma}}}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} = \prod_{i=1}^{n}\frac{1}{\zeta_i\sqrt{2\pi}}e^{-\frac{y_i^2}{2\zeta_i^2}} = \prod_{i=1}^{n} f_{\mathrm{N}}(y_i|0,\zeta_i^2). \tag{195}$$

In the new coordinate system the joint Gaussian probability distribution is expressed simply as the product of single-variate Gaussian marginals.

7.3.1.4 Marginal and Conditional Gaussian Distributions

Given the n-dimensional multivariate Gaussian distribution, $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we can partition the vector space in two subspaces, each of which is governed by a random distribution. If the partition aligns with the coordinates of the original random vector, then we can also partition the mean vector and covariance matrix, so that

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \Rightarrow \begin{cases} \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \\ \\ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \end{cases} \tag{196}$$

Following the discussion in §2.1.9 above, we show here that the joint distribution can be expressed as the product of marginal and conditional distributions, $\mathbf{X} = \mathbf{X}_2\,\mathbf{X}_1|(\mathbf{X}_2 = \mathbf{x}_2)$, all of which are Gaussian in form.

As a preliminary, we express the quadratic form that comprises the argument of the exponential function in block form,

$$(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu} \\ \mathbf{x}_2 - \boldsymbol{\mu} \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu} \\ \mathbf{x}_2 - \boldsymbol{\mu} \end{pmatrix} \tag{197}$$

using the upper Schur block (*cf. LAN*, §6.1) to decompose the block inverse into the product,

$$\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} I_p & 0 \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & I_q \end{pmatrix} \begin{pmatrix} \left(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)^{-1} & 0 \\ 0 & \boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix} \begin{pmatrix} I_p & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ 0 & I_q \end{pmatrix} \tag{198}$$

which can be expanded to yield

$$\begin{aligned} &(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \\ &= \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu} \\ \mathbf{x}_2 - \boldsymbol{\mu} \end{pmatrix}^\top \begin{pmatrix} I_p \\ \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \end{pmatrix} \left(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)^{-1} \begin{pmatrix} I_p \\ -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix}^\top \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu} \\ \mathbf{x}_2 - \boldsymbol{\mu} \end{pmatrix} + (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \end{aligned} \tag{199}$$

It is a simple matter to show that the multiplicative normalizing values also split up in a similar fashion so that the joint Gaussian probability density function factors into the product of Gaussians,

$$\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathrm{N}(\boldsymbol{\mu}_1 + (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})\ \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}). \tag{200}$$

This yields the desired result

$$\text{Marginal Distribution: } \mathbf{X}_2 \sim \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \tag{201}$$

$$\text{Conditional Distribution: } \mathbf{X}_1|(\mathbf{X}_2 = \mathbf{x}_2) \sim \mathrm{N}(\boldsymbol{\mu}_1 + (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}) \tag{202}$$

7.3.1.5 Mean and Variance of Sampled IID Normal Random Variables

An **IID normal random variable** is a random vector for which

- each coordinate is a normal random variable;
- the means and variances of the coordinate random variable are identical;
- the covariance of distinct coordinate random variables is identically zero:

$$\left.\begin{aligned} &\mathbf{X} = (X_1, \cdots, X_m)^\top, X_i \sim \mathrm{N}(\mu, \sigma^2) \\ &\mathbf{1}_m = (1, \cdots, 1)^\top \end{aligned}\right\} \Rightarrow \mathbf{X} \sim \prod_{i=1}^{m} X_i = \mathrm{N}(\mu \mathbf{1}_m, \sigma^2 I_m) \tag{203}$$

IID random variables are frequently used as models for sampling activities, for which the dimension of the space matches the number of sample points. IID *normal* random variables have the collective properties as summarized in **Fisher's Theorem**: the sample mean and sample variance – *cf.* §3.3.1 – are independent and distributed as Gaussian and chi-square random variables, respectively.

Sketch of Proof

- The mean and (unbiased) variance of the m coordinates of a random vector, $\mathbf{x} = (x_1, \cdots, x_m)^\top$, are defined as

$$\hat{\mu}\mathbf{1}_m = \mathbf{1}_m \frac{1}{m}\mathbf{1}_m^\top \mathbf{x} = \mathbf{1}_m(\mathbf{1}_m^\top \mathbf{1}_m)^{-1}\mathbf{1}_m^\top \mathbf{x} = P_{\mathbf{1}_m}\mathbf{x}; \tag{204}$$

$$\hat{\sigma}^2 = \frac{1}{m-1}(\mathbf{x} - \hat{\mu}\mathbf{1}_m)^\top(\mathbf{x} - \hat{\mu}\mathbf{1}_m) = \frac{1}{m-1}(\mathbf{x} - P_{\mathbf{1}_m}\mathbf{x})^\top(\mathbf{x} - P_{\mathbf{1}_m}\mathbf{x}) = \frac{1}{m-1}\mathbf{x}^\top(I_m - P_{\mathbf{1}_m})\mathbf{x}. \tag{205}$$

  Here, the vector that defines the mean and the collection of vectors that define the variance lie in orthogonal subspaces. The matrix operator, $P_{\mathbf{1}_m}$, projects vectors into the 1-dimensional subspace that contains the ones vector, while the matrix operator, $(I_m - P_{\mathbf{1}_m})$, projects vectors into the complementary orthogonal $(m-1)$-dimensional subspace.

- There exists an orthogonal basis for which one unit vector – say, the m^{th} one – is aligned with the sample mean vector, and the remainder lie within the complementary subspace. A change of basis to this coordinate system is achieved by a linear operator, the $m \times m$ matrix, $Q^\top$, so that the transformed random variable, $\mathbf{Y} = Q^\top \mathbf{X}$, takes the form

$$\mathbf{Y} = (Y_1, \cdots, Y_m)^\top = Q^\top \mathbf{X} \Rightarrow \mathbf{Y} \sim \mathrm{N}(\mu Q^\top \mathbf{1}_m, \sigma^2 Q^\top I_m Q) = \mathrm{N}(\sqrt{m}\,\mu \mathbf{e}_m, \sigma^2 I_m),$$
$$\mathbf{Y} = \prod_{i=1}^{m} Y_i \rightarrow \begin{cases} Y_1, \cdots, Y_{m-1} \sim \mathrm{N}(0, \sigma^2) \\ Y_m \sim \mathrm{N}(\sqrt{m}\,\mu, \sigma^2) \end{cases} \tag{206}$$

  From the presentation in §7.3.1.2 independence in multivariate Gaussian distributions is preserved under orthogonal linear transformation.

- The sample mean and variance are expressed more simply in the new coordinate system:

$$m\hat{\mu} \rightarrow \mathbf{1}_m^\top \mathbf{X} = \mathbf{1}_m^\top Q Q^\top \mathbf{X} = (Q^\top \mathbf{1}_m)^\top Q^\top \mathbf{X} = \sqrt{m}\,\mathbf{e}_m^\top \mathbf{Y} = \sqrt{m}\,Y_m \tag{207}$$

$$(m-1)\hat{\sigma}^2 \rightarrow \mathbf{X}^\top(I_m - P_{\mathbf{1}_m})\mathbf{X} = \mathbf{X}^\top Q Q^\top (I_m - P_{\mathbf{1}_m}) Q Q^\top \mathbf{X} = (Q^\top \mathbf{X})^\top (I_m - Q^\top P_{\mathbf{1}_m} Q)(Q^\top \mathbf{X})$$
$$= \mathbf{Y}^\top (I_m - \mathbf{e}_m \mathbf{e}_m^\top)\mathbf{Y} = \sum_{i=1}^{m-1} Y_i^2 \tag{208}$$

  from which the distributions of sample statistics can be immediately derived:

$$\text{sample mean: } \sqrt{m}\,\hat{\mu} = Y_m \sim \mathrm{N}\left(\mu, \sigma^2\right); \tag{209}$$

$$\text{sample variance: } (m-1)\frac{\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^{m-1} \frac{Y_i^2}{\sigma^2} = \sum_{i=1}^{m-1} Z_i^2 \sim \chi^2_{m-1}. \tag{210}$$

Notice that information distributed homogenously in the original coordinate system – all random variables, X_i, are identical – is distributed *inhomogeneously* in the transformed coordinate system in (206) – Y_m is distributed differently than $Y_1, \cdots, Y_{m-1}$. Since distinct coordinates of the transformed joint distribution contribute to the distributions of sample mean and sample variance, the two sample statistics are independent.

7.3.2 Wishart Distribution

The **Wishart** distribution is governed by random variables of *scatter matrices* formed from *data matrices* sampled from zero-mean multivariate Gaussian distributions. Wishart distributions are used to model sample covariance matrices.

Given the random data matrix, $\mathcal{X}$, formed by vectors drawn m times from a p-dimensional multivariate Gaussian distribution with mean zero and covariance matrix, Σ,

$$\mathcal{X} = \left(\mathbf{X}_1, \cdots, \mathbf{X}_m\right)^\top, \quad \mathbf{X}_i \sim \mathrm{N}(\mathbf{0}, \Sigma). \tag{211}$$

Here, the p-dimensional point is encoded in row vectors, while the matrix column vectors hold information on each variate. The associated random scatter matrix, $\mathcal{S} = \mathcal{X}^\top \mathcal{X}$, then follows a Wishart distribution,

$$\mathcal{S} \sim \mathrm{W}_p(\Sigma, m) \equiv f_{\mathrm{W}(\Sigma,m)}(S) = \frac{2^{-pm/2}\pi^{-p(p-1)/4}}{\prod_{i=1}^{p}\Gamma\left(\frac{m-i+1}{2}\right)} \frac{(\det S)^{(m-p-1)/2}}{(\det \Sigma)^{m/2}} e^{-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}S)}. \tag{212}$$

Here, the number of draws exceeds the dimension of the multivariate source, $m > p$.

To aid in the exposition in this section font and type are used to distinguish between the various quantities:

- ordinary majuscule fonts are used for sample matrices;
- boldface fonts are used for vectors: majuscule for random vectors, minuscule for sample vectors;
- calligraphic fonts are used for random matrices.

7.3.2.1 Sketch of Derivation

We start with the joint distribution of m independent draws from the p-dimensional Gaussian distribution,

$$\mathbf{X}_i \sim \mathrm{N}(\mathbf{0}, \Sigma) = \frac{1}{\sqrt{(2\pi)^p \det \Sigma}} e^{-\frac{1}{2}\mathbf{x}_i^\top \Sigma^{-1}\mathbf{x}_i} \Rightarrow$$

$$\mathcal{X} \sim \frac{(2\pi)^{-pm/2}}{(\det \Sigma)^{m/2}} \prod_{i=1}^{m} e^{-\frac{1}{2}\mathbf{x}_i^\top \Sigma^{-1}\mathbf{x}_i} = \frac{(2\pi)^{-pm/2}}{(\det \Sigma)^{m/2}} e^{-\frac{1}{2}\sum_{i=1}^{m}\mathbf{x}_i^\top \Sigma^{-1}\mathbf{x}_i}. \tag{213}$$

Given properties of the matrix trace operator, (*cf.* *LAN*, §3.10), the exponential argument can be expressed more succinctly in terms of matrices,

$$\sum_{i=1}^{m} \mathbf{x}_i^\top \Sigma^{-1}\mathbf{x}_i = \operatorname{tr}(X\Sigma^{-1}X^\top) = \operatorname{tr}(\Sigma^{-1}X^\top X), \tag{214}$$

and the probability density function takes the form

$$\mathcal{X} \sim \frac{1}{(2\pi)^{pm/2}(\det \Sigma)^{m/2}} e^{-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}X^\top X)}. \tag{215}$$

The derivation of the Wishart distribution then follows from the change of variable

$$X: \qquad m \times p \text{ variables} \tag{216}$$

$$S = X^\top X: \frac{p(p+1)}{2} \text{ variables} \tag{217}$$

and integrating the unconstrained variables so that, for expectation of an arbitrary function, g, we have

$$\int_{\mathbb{R}^{m\times p}} g(X) f_{\mathrm{N}(0,\Sigma)}(X)\, dX \to \int_{\mathbb{R}^{p(p+1)/2}} g(S) f_{\mathrm{W}(\Sigma,m)}(S)\, dS. \tag{218}$$

This is achieved by the following program:

- Define the sequence of matrices

$$X_i = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_i \end{pmatrix} \tag{219}$$

$$S_i = X_i^\top X_i = \begin{pmatrix} \mathbf{x}_1^\top \mathbf{x}_1 & \cdots & \mathbf{x}_1^\top \mathbf{x}_i \\ \vdots & \ddots & \vdots \\ \mathbf{x}_1^\top \mathbf{x}_i & \cdots & \mathbf{x}_i^\top \mathbf{x}_i \end{pmatrix} \tag{220}$$

The determinant of the i^{th} scatter matrix yields the square volume of the i-dimensional parallelepiped enclosed by the data vectors (*cf. LAN* , §3.7):

$$V_i^2 = \det(X_i^\top X_i) = \det(S_i); \tag{221}$$

- Map the column vectors, $\mathbf{x}_i$, into a set of orthogonal vectors, $\mathbf{u}_i$ that enclose equivalent volumes via the Gram-Schmidt procedure (*cf. LAN*, §3.11.1). The norm of the orthogonal vectors is given by

$$\left.\begin{array}{l} \mathbf{u}_1 = \mathbf{x}_1 \\ \quad \vdots \\ \mathbf{u}_i = \left(I - \sum_{j=1}^{i-1} P_{\mathbf{u}_j}\right) \mathbf{x}_i \end{array}\right\} \Rightarrow ||\mathbf{u}_i||_2 = \left(\frac{\det(X_i^\top X_i)}{\det(X_{i-1}^\top X_{i-1})}\right)^{\frac{1}{2}} = \left(\frac{\det(S_i)}{\det(S_{i-1})}\right)^{\frac{1}{2}} = \frac{V_i}{V_{i-1}} \tag{222}$$

- Partition the $m \times p$-dimensional space by variate, so that each m-dimensional vector is mapped into a smaller subspace spanned by the elements of the corresponding vector in the covariance matrix:

$$\left.\begin{array}{l} dX = \prod_{i=1}^p d\mathbf{x}_i \\ dS = \prod_{i=1}^p d\mathbf{s}_i \end{array}\right\} \Rightarrow \mathbf{x}_i = \begin{pmatrix} dx_{1i} \\ \vdots \\ dx_{mi} \end{pmatrix} \rightarrow \begin{pmatrix} ds_{1i} \\ \vdots \\ ds_{ii} \\ d\phi_{i+1,1} \\ \vdots \\ d\phi_{mi} \end{pmatrix} = \begin{pmatrix} d\mathbf{s}_i \\ d\boldsymbol{\phi}_i \end{pmatrix}; \tag{223}$$

- The m-dimensional vector, $\mathbf{u}_i$, is orthogonal to the $(i-1)$-dimensional subspace spanned by the data vectors, $(\mathbf{x}_1, \cdots, \mathbf{x}_{i-1})$, and is constrained by the i entries in the covariance vector, $\mathbf{s}_i = (\mathbf{x}_1^\top \mathbf{x}_i, \cdots, \mathbf{x}_i^\top \mathbf{x}_i)^\top$, to be constant on the $(m-i)$-dimensional hypersphere of radius, $||\mathbf{u}_i||_2$. We can therefore integrate the angular coordinates, $d\boldsymbol{\phi}_i$ for a given set of data vectors, $\mathbf{x}_1, \cdots, \mathbf{x}_{i-1}$,

$$\int d\boldsymbol{\phi}_i = \int d\phi_{i+1,i} \cdots d\phi_{mi} = A_{m-i}\left(||\mathbf{u}||_2\right) = A_{m-i}\left(\frac{V_i}{V_{i-1}}\right); \tag{224}$$

- The Jacobian of the transformation from data to scatter vectors is the reciprocal of the inverse transformation:

$$\nabla_{\mathbf{x}_i} \mathbf{s}_i = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_{i-1} & 2\mathbf{x}_i \end{pmatrix} \Rightarrow$$

$$J(\mathbf{x}_i, \mathbf{s}_i) = J^{-1}(\mathbf{s}_i, \mathbf{x}_i) = |\det\left(\nabla_{\mathbf{x}_i} \mathbf{s}_i^\top \nabla_{\mathbf{x}_i} \mathbf{s}_i\right)|^{-\frac{1}{2}} = |4\det(S_i)|^{-\frac{1}{2}} = \frac{1}{2V_i}. \tag{225}$$

The multiplicative factor in the Jacobian is due to the linearity of the wedge product from which the determinant is derived;

- Finally, apply Fubini's theorem so that

$$\int_{\mathbb{R}^{m\times p}} dX = \int_{\mathbb{R}^m} \left[\cdots \left[\int_{\mathbb{R}^m} d\mathbf{x}_p\right] \cdots\right] d\mathbf{x}_1 = \prod_{i=1}^p \frac{A_{m-i}\left(V_i/V_{i-1}\right)}{2V_i} \int_{\mathbb{R}^{p(p+1)/2}} dS. \tag{226}$$

Substituting the formula for the surface area of hyperspheres, supplying the exponential kernel to the integral, and replacing data matrices with the scatter matrix in (215) yields the form of the density function,

$$\left.\begin{aligned} A_{m-i}\left(\tfrac{V_i}{V_{i-1}}\right) &= \tfrac{2\pi^{(m-i+1)/2}}{\Gamma\left(\frac{m-i+1}{2}\right)}\left(\tfrac{V_i}{V_{i-1}}\right)^{m-i} \\ V_{m-p-1} &= (\det(S))^{\frac{1}{2}} \end{aligned}\right\} \Rightarrow f_{\mathrm{N}(0,\Sigma)}(X)\, dX \to f_{\mathrm{W}(\Sigma,m)}(S)\, dS, \tag{227}$$

as defined above in (212).

7.4 The Exponential Family of Distributions

A large number of the common parametrized distributions introduced above in §§7.1 – 7.3 can be expressed in factored exponential form, for which the contributions due to the random variable and the parameter are segregated into distinct functions. In its most-general form the single- and multivariate expressions for distributions in the **exponential family** are factored as,

$$\text{univariate: } p_X(x|\theta) = h(x)\exp(\eta(\theta)\,T(x) - A(\eta(\theta))), \tag{228}$$

$$\text{multivariate: } p_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})\exp(\eta(\boldsymbol{\theta})^{\top}T(\mathbf{x}) - A(\eta(\boldsymbol{\theta}))), \tag{229}$$

for which the random variable and parameter are coupled in the argument of the exponential through a product of functions. The function of the random variable, $T(X)$, is known as a *statistic*, which is defined and further discussed in *SN*, §2.2.1.

The function, $A(\eta(\theta))$, is known as the *log-partition function*, from which all statistical information about the distribution can be derived. Taking the univariate expression, for example, we can rearrange the terms in the probability density function,

$$\int_{\mathbb{R}} h(x)\exp(\eta(\theta)\,T(x) - A(\eta(\theta)))dx = 1 \Rightarrow A(\eta(\theta)) = \ln\left(\int_{\mathbb{R}} h(x)\exp(\eta(\theta)\,T(x))dx\right). \tag{230}$$

The moment-generating function for the statistic, $T(x)$, is the Laplace transform of the probability density function (*cf.* §5.1.1),

$$\begin{aligned} M_{T(x)}(u) = \mathbb{E}e^{uT(x)} &= \int_{\mathbb{R}} [\exp(uT(x))]h(x)\exp(\eta T(x) - A(\eta))dx = \int_{\mathbb{R}} h(x)\exp((\eta+u)T(x) - A(\eta))dx \\ &= \int_{\mathbb{R}} h(x)\exp[(\eta+u)T(x) - A(\eta+u) + A(\eta+u) - A(\eta)]\,dx \\ &= \exp(A(\eta+u) - A(\eta))\int_{\mathbb{R}} h(x)\exp[(\eta+u)T(x) - A(\eta+u)]\,dx \\ &= \exp(A(\eta+u) - A(\eta))\int_{\mathbb{R}} p_X(x|(\eta+u))\,dx = \exp(A(\eta+u) - A(\eta)) \end{aligned} \tag{231}$$

The cumulant function (*cf.* §5.1.3) is the natural logarithm of the moment-generating function, from which the first moment and second central moment are immediately derived:

$$K_{T(x)}(u) = \ln M_{T(x)}(u) = A(\eta+u) - A(\eta) \Rightarrow \begin{cases} \mathbb{E}T(x) = \frac{d}{du}\,K_{T(x)}(u)\big|_{u=0} &= \frac{d}{d\eta}A(\eta) \\ \mathbb{V}T(x) = \frac{d^2}{du^2}\,K_{T(x)}(u)\big|_{u=0} &= \frac{d^2}{d\eta^2}A(\eta) \end{cases} \tag{232}$$

Note that both moments are derivatives of the log-partition function, and is easily generalized to multivariate distributions:

$$\mathbb{E}T(\mathbf{x}) = \nabla_{\mathbf{u}}\,K_{T(\mathbf{x})}(\mathbf{u})\big|_{\mathbf{u}=\mathbf{0}} = \nabla_{\boldsymbol{\eta}}A(\boldsymbol{\eta}) \tag{233}$$

$$\mathbb{C}(T(\mathbf{x}),T(\mathbf{x})) = \nabla^2_{\mathbf{u}}\,K_{T(\mathbf{x})}(\mathbf{u})\big|_{\mathbf{u}=\mathbf{0}} = \nabla^2_{\boldsymbol{\eta}}A(\boldsymbol{\eta}) \tag{234}$$

Finally, the set of distributions for which the parameter function and statistics are both identity maps,

$$\left.\begin{array}{l}\eta(\theta) = \theta \\ T(x) = x\end{array}\right\} \Rightarrow p_X(x|\theta) = h(x)\exp(x\theta - A(\theta)), \tag{235}$$

is known as the **simple exponential family**.

8 Order Statistics

The probability distributions of cumulative probability rank, or 'percentile', of finite samples taken with replacement from *arbitrary* distributions are known as **order statistics**. The key insight is that *all* ranks are distributed uniformly with respect to their cumulative distribution function.

With this is mind we let a set of n IID random variables, designated as $X_1, \cdots, X_n$, be sampled from a uniform distribution, $X_i \sim \text{Uni}(0, 1)$. 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 $[u + du, 1]$.

Treating the infinitesimal interval as finite-sized, the frequency with which each independend sample falls within each interval is governed by multinomial statistics, §7.1.1.6,

$$\frac{n!}{(k-1)!1!(n-k)!}u^{k-1} \cdot du \cdot (1 - u - du)^{n-k} \approx \frac{n!}{(k-1)!(n-k)!}u^{k-1} \cdot (1-u)^{n-k} \cdot du \tag{236}$$

which, in the limit of infinitesimal interval, du, takes the form of a cumulative distribution whose probability density is given by

$$X_{(k)} \sim \text{B}(k, n + 1 - k). \tag{237}$$

9 Asymptotic Limits

9.1 Convergence of Random Variables

Random variables, X, as described in §2.1.5, map outcomes from the probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, into real values the state space, $(\mathbb{R}, \mathcal{B}, \mu)$, while the inverse map, X^{-1}, pulls back Borel sets from the state space into events in the probability space:

$$X : \Omega \to \mathbb{R} \tag{238}$$

$$X^{-1} : \mathcal{B} \to \mathcal{F} \tag{239}$$

Notice the difference in granularity of the forward and inverse map: the random variable applies to individual outcomes realized as a many-to-one mapping; the inverse action acts upon real sets, and cannot generally resolve individual outcomes.

Given a sequence of random variables, $X_{n\in\mathbb{N}}$, and a candidate limiting random variable, X, *convergence* is established by applying probability measures to a selection of events with vanishing results. However, the

mathematical setting is complex and convergence can be defined in a number of ways – one can initiate the selection of sets from either the probability space or the state space, or one can alter the sequence the limit and probability operations. More formally we can define convergence of random variables as the following:

- In the *inverse* sense as the selection of sets is initiated in the state space and the probability measure is applied to the sequence of events generated by the pullback: **convergence in distribution**;
- In the *forward* sense as the selection of set is initiated in the probability space, to which the probability measure is then applied; vanishing probability of the event in which the sequence and the limiting random variables differ:
  - **convergence in probability**: Application of the probability measure is made to each event, and convergence is demonstrated as the vanishing limit of real numbers;
  - **convergence almost surely**: Application of the probability measure is made to a limiting set, which requires that the limiting set be an event, and convergence is demonstrated as a zero-valued measure.

These three notions of convergence are covered briefly in the following sections. For the most part it is a simple matter of decoding the mathematical statements.

9.1.1 Convergence in Distribution

The cumulative distribution function, $F(x)$, completely defines the operation of random variables in the state space, and serves as the probability measure of half-open infinite sets, $[-\infty, x)$. Demonstration of convergence with respect to these sets is sufficient to establish convergence with respect to any Borel set. The pullback of the half-open set by the random variables, X_n, is represented as

$$X^{-1}[-\infty, x) = \{\omega \in \Omega : X(\omega) < x\} \equiv \{X < x\} \tag{240}$$

Convergence in distribution is established in a few steps:

- Generate the events that correspond to each random variable in the sequence: $\{X_n < x\}$
- Calculate the probability of each event: $\mathbb{P}\{X_n < x\}$
- Equate the limiting probability to the probability of the candidate random variable:

$$\lim_{n\to\infty} \mathbb{P}\{X_n < x\} = \mathbb{P}\{X < x\} \Leftrightarrow X_n \xrightarrow{d} X \tag{241}$$

Notice that the limit is applied to real values, and that the relation in (241) must hold for all x.

9.1.2 Convergence in Probability (Weak Convergence)

We start with the set of all outcomes for which the absolute difference between mappings, for an arbitrary random variable in the sequence, X_n, and the target random variable, X, exceeds a threshold, $\epsilon > 0$,

$$\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\} \equiv \{|X_n - X| > \epsilon\} \tag{242}$$

Then convergence in probability is established in a few steps:

- Generate the difference event for each random variable: $\{|X_n - X| > \epsilon\}$
- Calculate the probability of each event: $\mathbb{P}\{|X_n - X| > \epsilon\}$

- Assert the probability vanishes in the limit for all $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\{|X_n - X| > \epsilon\} = 0 \Leftrightarrow X_n \xrightarrow{p} X. \tag{243}$$

As with convergence in distribution, convergence in probability is a statement about the vanishing limit of real numbers.

9.1.3 Convergence Almost Surely (Strong Convergence)

Similar to convergence in probability, convergence almost surely starts with the sequence of difference sets defined in (242). However, the order of limit and measure operations is interchanged. The limit of a sequence of sets is best interpreted as the limit superior, defined as

$$\limsup_{n\to\infty}\{|X_n - X| > \epsilon\} = \bigcap_{n\in\mathbb{N}} \bigcup_{m\geq n} \{\omega \in \Omega : |X_m(\omega) - X(\omega)| > \epsilon\}. \tag{244}$$

The order of unions and intersections ensures that all outcomes that appear in an infinite number of sets in the sequence are assigned to the limiting set. Convergence almost surely then follows from applying the probability measure to the limiting set, and asserting the result vanishes for all $\epsilon > 0$,

$$\mathbb{P}\limsup_{n\to\infty}\{|X_n - X| > \epsilon\} = 0 \Leftrightarrow X_n \xrightarrow{as} X. \tag{245}$$

There is a subtlety here: the set of events, $\mathcal{F}$, must be complete, and contain the limiting set, which may not be the case!

9.1.4 Functions of Convergent Random Variables

The **Continuous Mapping Theorem**:

$$\left.\begin{array}{l} g \text{ is a } \textit{continuous} \text{ function} \\ X_n \xrightarrow{d} X \end{array}\right\} \Rightarrow g(X_n) \xrightarrow{d} g(X). \tag{246}$$

9.1.5 Product of Convergent Random Variables

Slutsky's Theorem:

$$\left.\begin{array}{l} X_n \xrightarrow{d} X \\ Y_n \xrightarrow{p} c \end{array}\right\} \Rightarrow X_n Y_n \xrightarrow{d} cX \tag{247}$$

Note that the convergence of the product holds *only* if the limit of one is a constant.

9.2 Asymptotic Limits of IID Samples

Given a random variable, X, with finite mean and variance,

$$\mathbb{E}X = \mu; \tag{248}$$

$$\mathbb{V}X = \sigma^2; \tag{249}$$

and a collection of independent, identically distributed random variables, $X_{i\in\mathbb{N}} \sim X$, the sample mean and variance of the collection are given by

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \Rightarrow \begin{cases} \mathbb{E}\bar{X}_n = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i = \mu; \\ \mathbb{V}\bar{X}_n = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}X_i = \frac{\sigma^2}{n}. \end{cases} \tag{250}$$

due to the linearity of the expectation operator. Clearly, information in the random variables is evermore tightly organized about the mean, μ, as the number of sample points increases, since the variance approaches zero in the limit. As described above in §§ 9.1.1 – 9.1.3, convergence of random variables has multiple interpretations and it is shown in the following sections that the asymptotic random variable for the sample mean converges in probability, almost surely and in distribution as specified in the weak and strong versions of the law of large numbers and the central limit theorem, respectively.

9.2.1 Law of Large Numbers

The *law of large numbers* may be expressed in two senses, both weak and strong, for which the asymptotic distribution of sample mean can be proved to converge to the mean,

$$\textbf{Weak Law of Large Numbers} \qquad \bar{X}_n \xrightarrow{p} \mu; \tag{251}$$

$$\textbf{Strong Law of Large Numbers} \qquad \bar{X}_n \xrightarrow{as} \mu. \tag{252}$$

The weak law can be shown as a simple consequence of Chebyshev's Inequality (see §4.2). Set $g(X) = |\bar{X}_n - \mu|$, then

$$\mathbb{P}\{|\bar{X}_n - \mu| \geq \epsilon\} = \mathbb{P}\{(\bar{X}_n - \mu)^2 \geq \epsilon^2\} \leq \frac{\sigma^2}{n\epsilon^2} \Rightarrow \lim_{n\to\infty}\mathbb{P}\{|\bar{X}_n - \mu| \geq \epsilon\} = 0, \tag{253}$$

which is the condition for convergence in probability.

It is also possible to use Chebyshev's Inequality to prove the strong law, but this requires use of tools from Lebesgue integration.

9.2.2 Central Limit Theorem

Whereas the law of large numbers assures us that the distribution of the sample mean approaches the population mean of the underlying distributions, the **central limit theorem** provides the limiting distribution itself!

9.2.2.1 Univariate Theorem Given a random variable, X, with mean and variance, μ and σ^2, respectively, and a collection of independent, identically distributed random variables, $X_{i\in\mathbb{N}} \sim X$, the scaled sample mean converges in distribution to the standard normal,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathrm{N}(0,1). \tag{254}$$

The characteristic equation, as described above in §5.1.2, provides a ready proof of the theorem. Given the definition of the sample mean, we construct the sum of scaled variables,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \Rightarrow \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \sum_{i=1}^{n}\frac{X_i - \mu}{\sigma\sqrt{n}} \Rightarrow \begin{cases} n\mathbb{E}\left(\frac{X_i-\mu}{\sigma\sqrt{n}}\right) = 0 \\ n\mathbb{E}\left(\frac{X_i-\mu}{\sigma\sqrt{n}}\right)^2 = 1, \end{cases} \tag{255}$$

along with the mean and variance of each contribution to the sum. The characteristic function for any one of the collection is given by

$$\phi_X(t) = 1 + it\mathbb{E}X - \frac{t^2}{2}\mathbb{E}X^2 + \cdots \Rightarrow \phi_{\frac{X_i-\mu}{\sigma\sqrt{n}}}(t) = 1 - \frac{t^2}{2n} + \mathcal{O}\left(n^{-\frac{3}{2}}\right) \tag{256}$$

and the full collection to leading order is given by,

$$\phi_{\sum_{i=1}^n \frac{X_i-\mu}{\sigma\sqrt{n}}}(t) = \prod_{i=1}^n \phi_{\frac{X_i-\mu}{\sigma\sqrt{n}}}(t) = \prod_{i=1}^n \left(1 - \frac{t^2}{2n} + \mathcal{O}\left(n^{-\frac{3}{2}}\right)\right). \tag{257}$$

Finally, noting that the limit yields an exponential function, whose form matches the characteristic function of the standard normal distribution,

$$\lim_{n\to\infty} \prod_{i=1}^n \left(1 - \frac{t^2}{n} + \mathcal{O}\left(n^{-\frac{3}{2}}\right)\right) = \lim_{n\to\infty} \left(1 - \frac{t^2}{n} + \mathcal{O}\left(n^{-\frac{3}{2}}\right)\right)^n = e^{-\frac{t^2}{2}} = \phi_{\mathrm{N}(0,1)}(t). \tag{258}$$

9.2.2.2 Multivariate Theorem

The multivariate version of the central limit theorem is a straightforward generalization of the univariate case. Given the mean vector and covariance matrix of a random variable, $\mathbf{X}$,

$$\mathbb{E}\mathbf{X} = \boldsymbol{\mu}; \tag{259}$$

$$\mathbb{V}\mathbf{X} = \mathbb{E}\mathbf{X}\mathbf{X}^\top - \mathbb{E}\mathbf{X}\,\mathbb{E}\mathbf{X}^\top = \boldsymbol{\Sigma}; \tag{260}$$

the asymptotic distribution of the sample mean approaches a normal distribution,

$$\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i \Rightarrow \sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{261}$$

whose covariance matrix matches the underlying distribution.

For any given *scalar* function of a multivariate random variable, $g(\mathbf{X})$, we can construct a related asymptotic distribution, via the **delta method**,

$$\left.\begin{array}{l} \mathbb{E}g(\mathbf{X}) = g(\boldsymbol{\mu}) \\ \mathbb{V}g(\mathbf{X}) \approx \nabla g(\boldsymbol{\mu})^\top \boldsymbol{\Sigma} \nabla g(\boldsymbol{\mu}) \end{array}\right\} \Rightarrow \sqrt{n}\left(g(\mathbf{X}_n) - g(\boldsymbol{\mu})\right) \xrightarrow{d} \mathrm{N}\left(0, \nabla g(\boldsymbol{\mu})^\top \boldsymbol{\Sigma} \nabla g(\boldsymbol{\mu})\right). \tag{262}$$

For any given *vector* function of a multivariate random variable, $\mathbf{g}(\mathbf{X})$, we can construct a related asymptotic distribution for which the gradient operator, ∇, in (262) is replaced by the Jacobian matrix operator, J,

$$\left.\begin{array}{l} \mathbb{E}\mathbf{g}(\mathbf{X}) = \mathbf{g}(\boldsymbol{\mu}) \\ \mathbb{V}\mathbf{g}(\mathbf{X}) \approx \left(J_{\boldsymbol{\mu}}\mathbf{g}(\boldsymbol{\mu})\right)^\top \boldsymbol{\Sigma} \left(J_{\boldsymbol{\mu}}\mathbf{g}(\boldsymbol{\mu})\right) \end{array}\right\} \Rightarrow \sqrt{n}\left(\mathbf{g}(\mathbf{X}_n) - \mathbf{g}(\boldsymbol{\mu})\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \left(J_{\boldsymbol{\mu}}\mathbf{g}(\boldsymbol{\mu})\right)^\top \boldsymbol{\Sigma} \left(J_{\boldsymbol{\mu}}\mathbf{g}(\boldsymbol{\mu})\right)\right). \tag{263}$$

Applying the delta method to the linear transformation, $\mathbf{g}(\mathbf{X}) = \Gamma\mathbf{X}$, yields

$$\left.\begin{array}{l} \mathbb{E}\mathbf{g}(\mathbf{X}) = \Gamma\boldsymbol{\mu} \\ \mathbb{V}\mathbf{g}(\mathbf{X}) = \Gamma^\top \boldsymbol{\Sigma} \Gamma \end{array}\right\} \Rightarrow \sqrt{n}\Gamma\left(\mathbf{X}_n - \boldsymbol{\mu}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \Gamma^\top \boldsymbol{\Sigma} \Gamma\right). \tag{264}$$

10 Likelihood Function and Information Measures

10.1 Thermodynamic Entropy

The origins of information theory lie in thermodynamics, in which the multinomial arrangements of identical particles – a **macrostate** – are assigned values that depend on the number of permutations –

each a **microstate**. For a given physical system the microstates are defined by distinct energy levels, while possible macrostates are constrained by the total energy. In particular the **physical entropy** of the dominant macrostate can be identified with the *logarithm* of the number of microstates, for which the linear extensibility of physical entropy is a consequence.

10.2 Shannon Information and Entropy

The link between physical entropy and probability theory for the **microcanonical** assembly described in §10.1 is furnished by the equivalence between indistinguishability of particles and the equal probability of permutations. For a physical system of interacting particles, the macrostate that maximizes entropy – the one with the greatest number of microstates – is interpreted as the *most probable* to be observed at any given time. And since energy is exchanged during interactions, the instantaneous macrostates fluctuate about the maximum-entropy state, with the size of the excursions an inverse function of the number of particles in the multinomial permutations.

As the number of particles increases, the observations of macroscopic phenomena are dominated by a *single* most-probable macrostate, which becomes sharp in the limit. Given a random variable, X, the limiting microcanonical entropy is identified with the **Shannon information**,

$$\mathbb{I}_S X = -\ln \mathbb{P}_X. \tag{265}$$

For distinct mappings of the random variable the Shannon information varies by the logarithm of the probability measure, which can be thought of as due to the indistinguishability – and equal 'probability' – of the underlying outcomes that are collected in the density, $\mathbb{P}_X$. Since information can be linearly combined, the *global* information, or **Shannon entropy**, is captured by the expectation of Shannon information,

$$\mathbb{H}X \equiv \mathbb{E}\mathbb{I}_S X = \int_{-\infty}^{\infty} p_X(x) \ln p_X \, dx. \tag{266}$$

An alternative interpretation of information and entropy comes from coding theory, in which discrete random variables are efficiently encoded in bits. Here, information in the multinomial arrangements of bits is provided by the number required to represent the ensemble, which is provided by the logarithm.

A common application of Shannon entropy is to estimate the form of a distribution consistent with a finite set of measurements. The basic idea is that the 'true' distribution is one that is both

- consistent with the measurements;
- evenly distributed (maximizes 'disorder') otherwise.

10.3 Joint Entropy Measures

There are a number of quantities derived from information and entropy that measure the relation between two different random variables, which are taken up in the next few sections.

10.3.1 Relative Entropy (Kullback-Leibler Divergence)

The **relative entropy** between random variables, X and Y, is given by

$$D_{\mathrm{KL}}(X||Y) \equiv \mathbb{H}X - \mathbb{H}_X Y = -\int_{-\infty}^{\infty} p_X(x) \ln \frac{p_X(x)}{p_Y(x)} \, dx. \tag{267}$$

The relative entropy is defined as the difference between the entropy of the random variable, X, and the expection with respect to X of the information in Y. If the two random variables coincide, the relative entropy is zero; otherwise, by Jensen's Inequality, the relative entropy must be positive:

$$D_{\mathrm{KL}}(X||Y) = -\int_{-\infty}^{\infty} p_X(x) \ln \frac{p_X(x)}{p_Y(x)}\, dx \geq -\ln\left(\int_{-\infty}^{\infty} p_X(x) \frac{p_X(x)}{p_Y(x)}\, dx\right) = -\ln \int_{-\infty}^{\infty} p_Y(x)\, dx = 0. \tag{268}$$

The relative entropy is interpreted as a measure of the error introduced by substituting the random variable, Y, for cases in which X is correct. Notice also that the relative entropy is *not* symmetric in X and Y, and the relative entropy of Y with respect to X is generally different for the relative entropy of X with respect to Y.

10.3.2 Conditional Entropy

The **conditional entropy** the random variable, $X|Y$, is provided by the linear combination of the entropy of the joint distribution less the entropy in the conditional variable,

$$\mathbb{H}(X|Y) = \mathbb{H}(X,Y) - \mathbb{H}Y. \tag{269}$$

10.3.3 Mutual Information

The **mutual information** between the random variables, X and Y, can be expressed either in terms of entropy or in terms of the Kullback-Leibler divergence,

$$\mathbb{I}_M(X,Y) = \mathbb{H}X + \mathbb{H}Y - \mathbb{H}(X,Y) = D_{\mathrm{KL}}\left((X,Y)||\, XY\right) = \mathbb{E}_Y D_{\mathrm{KL}}\left(X|Y||\, X\right). \tag{270}$$

Notice that, unlike relative entropy, mutual information is symmetric with respect to the random variables that comprise the joint distribution. Mutual information measures the 'overlap' entropy between the two variables that make up the joint distribution. For independent random variables the mutual information is the sum of the entropy in each.

10.4 Likelihood Function and Fisher Information

In the definitions of Shannon information and entropy, in (265) and (266), respectively, it is clear that information is a local property – a value defined at a single point in the domain – while entropy is a global one that summarizes the full distribution. A family of random variables defined over a parameter has, in some sense, local information encoded not only in the density function, but also in the change of densities due to change in the parameter.

Let the IID random vector, $\mathbf{X} = (X_1, \cdots, X_n)$, model the sampling distribution as described above in §2.1.8, with the additional requirement that the coordinate random variables be governed by a parameter, $X_i \sim X|\theta$. The joint probability distribution for the sample is given by the function, $p_{\mathbf{X}}(\mathbf{x}|\theta)$, and the **likelihood function**, $L(\theta|\mathbf{x})$, given that $\mathbf{X} = \mathbf{x}$ is observed, is defined as

$$L(\theta|\mathbf{x}) \equiv p_{\mathbf{X}}(\mathbf{x}|\theta) \tag{271}$$

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

$$\ln L(\theta|\mathbf{x}) = \ln p_{\mathbf{X}}(\mathbf{x}|\theta) = \ln \mathbb{P}\mathbf{X}_\theta = -\mathbb{I}_S \mathbf{X}_\theta. \tag{272}$$

Since the samples are independent, the Shannon information is additive, equal to the sum of information in each measurement,

$$L(\theta|\mathbf{x}) \equiv p_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} p_X(x_i|\theta) \Rightarrow \ln L(\theta|\mathbf{x}) = \sum_{i=1}^{n} \ln p_X(x_i|\theta). \tag{273}$$

10.4.1 The Distinction between Probability and Likelihood

Although similar in appearance, the likelihood function, $L(\theta|\mathbf{x})$, is *not* a probability distribution for θ – it is rather, from the definition in (271), equivalent to the distribution of the sample, $\mathbf{x}$. Take, for example, the binomial distribution described in §7.1.1.1,

$$\mathrm{Bin}(n,\theta) \rightarrow p_{\mathrm{Bin}}(k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \tag{274}$$

which is a discrete distribution for the number of successes, k, expected from n draws, for which each draw has a probability of success, θ. The likelihood function,

$$L(\theta|n,k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \tag{275}$$

is *not* a probability function for the parameter. There *is* a continuous probability function in θ with a similar form – this is the Beta distribution described in §7.2.3.1,

$$\mathrm{B}(\alpha,\beta) \Rightarrow p_{\mathrm{B}}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}. \tag{276}$$

We can align the forms of the two probability distributions by identifying parameters, using the combinatorial-Beta relation in (94), and expressing both probability distributions in terms of the parameters, k and n,

$$\left.\begin{array}{l}\alpha = k+1\\ \beta = n-k+1\end{array}\right\} \Rightarrow \begin{cases}\text{probability in successes:} & p_{\mathrm{Bin}}(k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}\\ \text{probability in parameter:} & p_{\mathrm{B}}(\theta|k+1,n-k+1) = (n+1)\binom{n}{k}\theta^k(1-\theta)^{n-k}\end{cases} \tag{277}$$

which makes clear the distinction between the two. The coincidence in form between probability and likelihood is addressed more fully with the Bayesian approach in §11, which partly relies on properties of the expontial family of distributions (*cf.* §7.4).

10.4.2 Fisher Information

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

$$S(\theta|\mathbf{x}) = \frac{\partial}{\partial\theta}\ln L(\theta|\mathbf{x}) = \frac{\partial}{\partial\theta}\ln p(\mathbf{x}|\theta) = \frac{1}{p(\mathbf{x}|\theta)}\frac{\partial}{\partial\theta}p(\mathbf{x}|\theta), \tag{278}$$

which is the derivative with respect to the parameter of the Shannon information in the joint distribution.

Given the two integration identities,

$$\int_D p\frac{\partial}{\partial\theta}\ln p\,d\mathbf{x} = \int_D \frac{\partial}{\partial\theta}p\,d\mathbf{x} = \frac{\partial}{\partial\theta}\int_D p\,d\mathbf{x} = 0 \tag{279}$$

$$\int_D p\left(\frac{\partial}{\partial\theta}\ln p\right)^2 d\mathbf{x} = \int_D \frac{1}{p}\left(\frac{\partial p}{\partial\theta}\right)^2 d\mathbf{x} = \int_D \left(\frac{\partial^2 p}{\partial\theta^2} - p\frac{\partial^2}{\partial\theta^2}\ln p\right) d\mathbf{x} = -\int_D p\frac{\partial^2}{\partial\theta^2}\ln p\,d\mathbf{x} \tag{280}$$

we can calculate the expection and variance of the score function:

$$\mathbb{E}_\theta S(\theta|\mathbf{X}) = 0 \tag{281}$$

$$\mathbb{V}_\theta S(\theta|\mathbf{X}) = \mathbb{E}_\theta S(\theta|\mathbf{X})^2 - \left(\mathbb{E}_\theta S(\theta|\mathbf{X})\right)^2 = \mathbb{E}_\theta S(\theta|\mathbf{X})^2 = -\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}S(\theta|\mathbf{X})\right) \tag{282}$$

Note that the expection of the score function, which is a kind of global property assigned point-wise, vanishes. The variance of the score function is known as **Fisher information**, defined formally as

$$\mathbb{I}_F\mathbf{X}_\theta = \mathbb{V}S(\theta|\mathbf{X}) = -\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}S(\theta|\mathbf{X})\right) = -\mathbb{E}_\theta\frac{\partial^2}{\partial\theta^2}\ln p(\mathbf{X}|\theta). \tag{283}$$

Therefore, the Fisher information describes the (local) change in (global) Shannon entropy as the parameter undergoes an infinitesimal change. Expanding the joint log-likelihood function into a power series in the parameter, via Taylor's theorem, yields

$$\ln L(\theta+\Delta\theta|\mathbf{x}) \approx \ln L(\theta|\mathbf{x}) + \frac{\partial}{\partial\theta}\ln L(\theta|\mathbf{x})\Delta\theta + \frac{1}{2}\frac{\partial^2}{\partial\theta^2}\ln L(\theta|\mathbf{x})\,\Delta\theta^2, \tag{284}$$

and applying the expectation operator to the result,

$$\begin{aligned}
-\mathbb{E}_\theta \ln L(\theta+\Delta\theta|\mathbf{X}) &\approx -\mathbb{E}_\theta\left(\ln L(\theta|\mathbf{X}) + \frac{\partial}{\partial\theta}\ln L(\theta|\mathbf{X})\Delta\theta + \frac{1}{2}\frac{\partial^2}{\partial\theta^2}\ln L(\theta|\mathbf{X})\Delta\theta^2\right) \\
&= \mathbb{E}_\theta\mathbb{I}_S\mathbf{X}_\theta - \mathbb{E}_\theta S(\theta|\mathbf{X})\Delta\theta + \frac{1}{2}\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}S(\theta|\mathbf{X})\right)\Delta\theta^2 \\
\mathbb{H}\mathbf{X}_{\theta+\Delta\theta} &\approx \mathbb{H}\mathbf{X}_\theta + \frac{1}{2}\mathbb{I}_F\mathbf{X}_\theta\Delta\theta^2
\end{aligned} \tag{285}$$

shows that Fisher information provides the estimate of the second-order global change in entropy given changes in the local parameter value.

A quick calculation shows that this is identical to the Kullback-Leibler divergence applied to nearby values of the parameter, θ,

$$\begin{aligned}
D_{\mathrm{KL}}\left(L(\theta|\mathbf{x})||L(\theta+\Delta\theta|\mathbf{x})\right) &= \int_\Theta L(\theta|\mathbf{x})\ln\frac{L(\theta|\mathbf{x})}{L(\theta+\Delta\theta|\mathbf{x})}\,d\theta \\
&= \int_\Theta L(\theta|\mathbf{x})\left(\ln L(\theta|\mathbf{x}) - \ln L(\theta+\Delta\theta|\mathbf{x})\right)\,d\theta \\
&\approx \int_\Theta L(\theta|\mathbf{x})\left(-\frac{\partial}{\partial\theta}\ln L(\theta|\mathbf{x})\Delta\theta - \frac{1}{2}\frac{\partial^2}{\partial\theta^2}\ln L(\theta|\mathbf{x})\Delta\theta^2\right)\,d\theta \\
&= -\mathbb{E}_\theta S(\theta|\mathbf{X})\,\Delta\theta - \frac{1}{2}\mathbb{E}_\theta\frac{\partial}{\partial\theta}S(\theta|\mathbf{X})\,\Delta\theta^2 \\
&= \frac{1}{2}\mathbb{I}_F\mathbf{X}_\theta\,\Delta\theta^2
\end{aligned} \tag{286}$$

and the Taylor expansion of log-likelihood can be equivalently expressed as

$$\mathbb{H}\mathbf{X}_{\theta+\Delta\theta} \approx \mathbb{H}\mathbf{X}_{\theta} + D_{\mathrm{KL}}\left(L(\theta|\mathbf{x})||L(\theta+\Delta\theta|\mathbf{x})\right). \tag{287}$$

10.4.3 Multidimensional Fisher Information

The ideas on Fisher information in §10.4.2 are readily extended to the case for which the parameters are multidimensional, expressed as the vector, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^\top$, and the joint random variable expressed as $\mathbf{X}_{\boldsymbol{\theta}}$. Here, the **Fisher information matrix** is given by

$$\mathbb{I}_F\mathbf{X}_{\boldsymbol{\theta}} \equiv -\mathbb{E}_{\boldsymbol{\theta}}\nabla^2 \ln p(\mathbf{X}|\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\begin{pmatrix} \frac{\partial^2}{\partial\theta_1^2}\ln p(\mathbf{X}|\boldsymbol{\theta}) & \cdots & \frac{\partial^2}{\partial\theta_1\partial\theta_m}\ln p(\mathbf{X}|\boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\theta_m\partial\theta_1}\ln p(\mathbf{X}|\boldsymbol{\theta}) & \cdots & \frac{\partial^2}{\partial\theta_m^2}\ln p(\mathbf{X}|\boldsymbol{\theta}) \end{pmatrix} \tag{288}$$

for which the univariate second-order derivative in (283) is replaced by the Laplacian operator, and the scalar result replaced by a matrix.

Also, as in the univariate case, the multidimensional Taylor expansion of the log-likelihood function,

$$\ln L(\boldsymbol{\theta}+\Delta\boldsymbol{\theta}|\mathbf{x}) \approx \ln L(\boldsymbol{\theta}|\mathbf{x}) + \Delta\boldsymbol{\theta}^\top \ln L(\boldsymbol{\theta}|\mathbf{x}) + \frac{1}{2}\Delta\boldsymbol{\theta}^\top H \ln L(\boldsymbol{\theta}|\mathbf{x})\Delta\boldsymbol{\theta} \tag{289}$$

leads to an identical relation and interpretation between Shannon entropy and Fisher information,

$$\mathbb{H}\mathbf{X}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}} \approx \mathbb{H}\mathbf{X}_{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta}^\top\frac{1}{2}\mathbb{I}_F\mathbf{X}_{\theta}\Delta\boldsymbol{\theta}. \tag{290}$$

And again, as in the univariate case, the Kullback-Leibler divergence,

$$D_{\mathrm{KL}}\left(L(\boldsymbol{\theta}|\mathbf{x})||L(\boldsymbol{\theta}+\Delta\boldsymbol{\theta}|\mathbf{x})\right) = \Delta\boldsymbol{\theta}^\top\frac{1}{2}\mathbb{I}_F\mathbf{X}_{\boldsymbol{\theta}}\,\Delta\boldsymbol{\theta} \tag{291}$$

supplies the value of the second-order correction,

$$\mathbb{H}\mathbf{X}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}} \approx \mathbb{H}\mathbf{X}_{\boldsymbol{\theta}} + D_{\mathrm{KL}}\left(L(\boldsymbol{\theta}|\mathbf{x})||L(\boldsymbol{\theta}+\Delta\boldsymbol{\theta}|\mathbf{x})\right). \tag{292}$$

10.4.4 Statistical Applications of Likelihood and Fisher Information

Parameter estimation is a major application for likelihood functions and Fisher information. Here, the problem is to estimate the parameter, θ, that governs a random variable, $X|\theta$, from a finite set of measurements, $x_1, \cdots, x_n$, each independently sampled from the distribution, $X_i \sim X \equiv X|\theta$. Roughly speaking, the parameter affects the likelihood of measurements, and the best estimate can be gained by determining the parameter value that maximizes the formal likelihood defined in (271) or, more typically, the log-likelihood defined in (272). Since the log-likelihood function is identical to the entropy for the parametrized distribution, the estimate generated by maximizing likelihood is equivalent one generated by maximizing entropy.

In the presence of ever-increasing information it is possible to show that the maximum-likelihood estimator described above is asymptotically normal, approaching the true value of the parameter with a covariance defined by the Fisher information, or the Fisher information matrix in the case of multidimensional distributions. Parameter estimation is a problem more naturally discussed in statistics than in probability proper, and is covered in the companion presentation, *Statistics Notes*, §3.3.

11 Bayesian Perspectives

There are two main interpretations of probabilistic modeling, and in particular determining the parameter, θ, that governs a model for a physical process:

- **the frequentist perspective:** the parameter, θ, is best interpreted as a property of the physical process. The parameter determines the frequency with which events occur, and the best estimate for its value is determined by measurement and a global framework for evaluating the expectation of frequencies given possible parameter values;
- **the Bayesian perspective:** the parameter, θ, is best interpreted as a property of measurement. The parameter is provisional and evolving, encoding a belief given prior measurement and a posterior updates given incoming information.

Clearly, the results converge asymptotically with the incorporation of infinite information. The transient differences – those realized by incorporation of finite information – are due to the choice of global framework and to the selection of prior and posterior parametric distributions. For specific choices of each, for example a maximum-likelihood global framework and uniform prior, the differences may vanish. In the absence of a verifiably preferred form for prior distribution the main advantages afforded the Bayesian perspective are computational convenience and efficiency.

11.1 Bayes' Theorem

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

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{B|A\}\mathbb{P}\{A\}}{\mathbb{P}\{B\}}, \tag{293}$$

which provides a means by which the event conditioning may be reversed. The statement in (293) can be interpreted either as a property of events in discrete distributions – for which the probabilities of the events, A and B, are determined by counting – or as continuous random variables, for which the theorem is expressed in terms of probability densities as in (30).

11.2 Prior, Posterior and Likelihood

Bayes' Theorem is frequently applied to the problem of parameter estimation introduced above in §10.4.4: estimate the parameter, θ, that governs a random variable, $X|\theta$, from a finite set of measurements, $x_1, \cdots, x_n$, each independently sampled from the distribution, $X_i \sim X \equiv X|\theta$. Whereas the maximum-likelihood or maximum-entropy approach is a straightforward maximization of the likelihood function, the Bayesian method raises the status of the parameter to a full-fledged random variable, and seeks solutions within the *joint* distribution, X, θ. In this way we can leverage Bayes' theorem, especially expressed as factors and products of probability densities as in (30), and develop an iterative method in terms of linked random variables, one that defines the uncertainty in the parameter, and a second that accounts for likelihood in the sampling process:

$\boldsymbol{\theta}_i$	the prior distribution for the parameter	(294)
$\boldsymbol{\theta}_{i+1}\|\mathbf{X}_i$	the posterior distribution for the parameter given measurements	(295)
$\mathbf{X_i}\|\boldsymbol{\theta_i}$	measurement distribution given prior parameter	(296)
$\mathbf{X_i}$	unconditioned measurement distribution	(297)

The alignment of the random variables with the events – A is the event that the parameter takes specific values, B is the event that specific measurements are realized – leads to the equivalent expression,

$$p(\theta_{i+1}) \equiv p(\theta_{i+1}|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|\theta_i)\, p(\theta_i)}{p(\mathbf{x}_i)} = \frac{p(\mathbf{x}_i|\theta_i)\, p(\theta_i)}{\int_\Theta p(\mathbf{x}_i|\theta_i')\, d\theta_i'}. \tag{298}$$

Again, a fuller justification within the topic of parameter estimation – in particular the grounding of the iterative expression in (298) in terms of minimizing **Bayes' risk** – is presented in a companion writeup, *Statistics Notes*. It is worth noting, however, that the maximum-likelihood method (usually presented as 'frequentist') is contained within the more general Bayesian approach given a uniform prior distribution for the parameter in the absence of any information afforded by measurement. Both can be expressed in iterative form. Within the more general Bayesian approach it is also possible to specify alternative priors with greater concentration of weight in regions of the parameter space judged more likely to contain the true value, and so develop a more efficient estimator in the event that the judgment is correct. Asymptotically, both maximum-likelihood and more general Bayesian approaches approach zero-mean Gaussian distributions whose covariance is proportional to the Fisher information contained within the measurements, $\mathbf{x}$.

11.3 Conjugate Families

The form of the prior and posterior distributions in the iterative form of Bayes's theorem, expressed above in (298), depend partly on the form selected for the likelihood function for measurements. For some choices of likelihood and prior, however, the posterior distribution matches the prior distribution *exactly in form*. These pairs are termed **conjugate families**, and reduce the calculations necessary to carry out the iterative program to simple arithmetic operations.

A number of common conjugate families are provided below, but in all cases both likelihood function and parameter density distributions largely coincide in form, with a kind of dual interchange between the active 'variables' and the supporting 'parameters'. In these cases the posterior distributions are generated by the product of likelihood and prior, normalized by the marginal for the prior. The form of numerator, however, identifies the posterior as another member of the same family as the prior, and the explicit calculation of the normalizing factor is unnecessary.

11.3.1 Exponential Family Prior and Likelihoods

Prior distribution:

$$H_n \sim f_H(\eta|\tau, n) \propto \exp(\tau\eta(\theta) - nA(\eta(\theta))) \tag{299}$$

Likelihood function:

$$X \sim f_X(x|\theta) = h(x)\exp(\eta(\theta)\, T(x) - A(\eta(\theta))) \Rightarrow$$
$$\mathbf{X} = (X_1, \cdots, X_m)^\top \sim f_{\mathbf{X}}(\mathbf{x}|\theta) = \left(\prod_{i=1}^{m} h(x_i)\right) \exp\left(\eta(\theta) \sum_{i=1}^{m} T(x_i) - mA(\eta(\theta))\right) \tag{300}$$

Posterior distribution:

$$H_{n+m} \sim f_H\left(\eta \,\middle|\, \tau + \sum_{i=1}^{m} T(x_i)\ , m+n\right) \propto \exp\left(\left(\tau + \sum_{i=1}^{m} T(x_i)\right)\eta(\theta) - (m+n)A(\eta(\theta))\right) \tag{301}$$

11.3.2 Beta Prior, Binomial Likelihood

$$\left.\begin{array}{ll} \text{beta prior:} & \mathrm{B}(\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ \text{binomial likelihood:} & \mathrm{Bin}(n,k) \propto \theta^{k}(1-\theta)^{n-k} \end{array}\right\} \Rightarrow \text{beta posterior: } \mathrm{B}(\alpha+k, \beta+n-k) \tag{302}$$

11.3.3 Gamma Prior, Posson Likelihood

$$\left.\begin{array}{ll}\text{gamma prior:} & \Gamma(\alpha,\beta) \propto x^{\alpha-1}e^{-\beta x} \\ \text{Poisson likelihood:} & \text{Poi}(x) \propto x^k e^{-nx}\end{array}\right\} \Rightarrow \text{gamma posterior: } \Gamma(\alpha+k,\beta+n) \tag{303}$$

11.3.4 Gaussian Prior, Gaussian Likelihood

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

$$\Rightarrow \text{Gaussian posterior: } \mu \sim \text{N}\left(\frac{\kappa_0\mu_0 + n\bar{\mathbf{x}}}{\kappa_0+n}, \frac{\sigma^2}{\kappa_0+n}\right) \tag{304}$$

$$\kappa_0 = \frac{\sigma}{\sigma_0} \tag{305}$$

11.3.5 Inverted Gamma Prior, Gaussian Likelihood

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

$$\Rightarrow \text{inverted gamma posterior: } \sigma^2 \sim \text{I}\Gamma\left(\alpha+\frac{n}{2}, \beta+\frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2\right) \tag{306}$$