Probability Notes

Mark DiBattista

May 2, 2019

Abstract

Probability.

1 Suggested Resource Materials

Useful source texts:

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

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

2 Probability Preliminaries

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

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

2.1 Probability Spaces, State Spaces and Random Variables

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

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

2.1.1 Set Collections

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

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

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

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

since the complementation of countable union is countable intersection.

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

2.1.2 Set Functions

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

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

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

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

which provides an alternative basis for the definition.

2.1.3 Measurability of Sets and Functions

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

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

and

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

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

2.1.4 Probability and State Spaces

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

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

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

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

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

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

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

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

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

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

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

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

2.1.5 Random Variables

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

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

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

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

2.1.6 The Probability Density and Cumulative Distribution Functions

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

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

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

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

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

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

2.1.7 The Expectation Operator

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

$$\mathbb{E}gX = \int_{-\infty}^{\infty} g(x)\,p_X(x)\,dx. \tag{18}$$

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

$$1_A(\omega) = \begin{cases} 1, \omega \in A \\ 0, \omega \notin A \end{cases} \quad \Rightarrow \mathbb{P}A = \mathbb{E}\,1_A, \tag{19}$$

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,

$$\mathbb{E}X^n = \int_{-\infty}^{\infty} x^n\,p_X(x)\,dx. \tag{20}$$

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

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

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,

$$\mathbb{E}(X - \mathbb{E}X)^n = \int_{-\infty}^{\infty} (x - \mu_X)^n\, p_X(x)\, dx. \tag{22}$$

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

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

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

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

3.2 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:

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

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

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

$$\text{sample mean:} \quad \bar{x} = \frac{1}{m}\sum_{x=1}^{m} x_i \tag{28}$$

$$\text{sample variance:} \quad s_x = \frac{1}{m}\sum_{x=1}^{m}(x_i - \bar{x})^2 = \frac{1}{m}\sum_{x=1}^{m} x_i^2 - \bar{x}^2. \tag{29}$$

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,

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix} = \begin{pmatrix} \leftarrow & \mathbf{x}_1 & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_m & \rightarrow \end{pmatrix} \equiv \begin{pmatrix} \uparrow & & \uparrow \\ \mathbf{c}_1 & \cdots & \mathbf{c}_n \\ \downarrow & & \downarrow \end{pmatrix}. \tag{30}$$

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 (28) and (29), and defining the m-dimensional ones vector, $\mathbf{1}_m = \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}$, we can express the *coordinate-wise* sample means and variances in terms of inner products (see *LAN*, §3.2),

$$\text{coordinate sample mean:} \quad \bar{x}_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij} = \frac{1}{m}\langle \mathbf{c}_i, \mathbf{1}_m \rangle \tag{31}$$

$$\text{coordinate sample covariance:} \quad s_{jk} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) = \frac{1}{m}\langle \mathbf{c}_i, \mathbf{c}_j \rangle - \bar{x}_i\bar{x}_j \tag{32}$$

Given the following matrix operations,

$$X^\top \mathbf{1}_m = m \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_n \end{pmatrix}, \tag{33}$$

$$X^\top \mathbf{1}_m (\mathbf{1}_m^\top \mathbf{1}_m)^{-1} \mathbf{1}_m^\top X = X^\top P_{\mathbf{1}_m} X = m \begin{pmatrix} \bar{x}_1^2 & \cdots & \bar{x}_1\bar{x}_n \\ \vdots & \ddots & \vdots \\ \bar{x}_n\bar{x}_1 & \cdots & \bar{x}_n^2 \end{pmatrix}, \tag{34}$$

for which the matrix, $P_{\mathbf{1}_m}$, is a projection matrix (see *LAN*, §3.10). The sample mean (vector) and sample variance (covariance matrix) can be expressed in vector and matrix form,

$$\text{sample mean vector:} \quad \bar{\mathbf{x}} = \frac{1}{m} X^\top \mathbf{1}_m \tag{35}$$

$$\text{sample covariance matrix:} \quad S_X = \frac{1}{m} X^\top X - \frac{1}{m} X^\top P_{\mathbf{1}_m} X = \frac{1}{m} X^\top (I_m - P_{\mathbf{1}_m}) X. \tag{36}$$

Notice that the information in the sample mean and sample covariance matrix lie in $P_{\mathbf{1}_m}$ and $(I_m - P_{\mathbf{1}_m})$, orthogonal 1- and $(n-1)$-dimensional subspaces, respectively.

4 Joint Distributions and Independence

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

$$p_{X,Y}(x, y) = p_X(x)p_Y(y). \tag{37}$$

This definition can be extended to cover multivariate distribution of arbitrary random variables, represented in vector form as $\mathbf{X} = \begin{pmatrix} X_1 & \cdots & X_n \end{pmatrix}$, with independence defined through

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} p_{X_i}(x_i). \tag{38}$$

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,

$$X_1, \ldots, X_n \sim X \Rightarrow p_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} p_X(x_i) \tag{39}$$

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

$$\left.\begin{array}{ll} \text{joint distribution:} & p_{X,Y}(x,y) \\ \text{marginal distribution:} & p_Y(y) \\ \text{conditional distribution:} & p_{X|Y}(x|y) \end{array}\right\} \Rightarrow p_{X,Y}(x,y) = p_{X|Y}(x|y)\, p_Y(y). \tag{40}$$

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

$$p_{X|Y} = \frac{p_{X,Y}(x,y)}{p_Y(y)}. \tag{41}$$

5 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.

5.1 Cauchy-Schwarz Inequality

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

$$|\langle \mathbf{x}, \mathbf{y} \rangle|^2 \leq \langle \mathbf{x}, \mathbf{x} \rangle \cdot \langle \mathbf{y}, \mathbf{y} \rangle. \tag{42}$$

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

$$\langle X, Y \rangle \equiv \mathbb{E}XY, \tag{43}$$

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,

$$|\mathbb{E}XY|^2 \leq \mathbb{E}X^2 \cdot \mathbb{E}Y^2 \Rightarrow \mathbb{C}(X,Y) \leq \mathbb{V}X \cdot \mathbb{V}Y \tag{44}$$

5.2 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 functions of the random variable, $g(X)$, via **Chebyshev's Inequality**,

$$\mathbb{P}\{g(X) \geq r\} \leq \frac{\mathbb{E}g(X)}{r}. \tag{45}$$

This follows directly from simple properties of integrals, since

$$\mathbb{E}g(X) = \int_D g(x)f(x)\,dx \geq \int_{\{g(x)\geq r\}} g(x)f(x)\,dx \geq r\int_{\{g(x)\geq r\}} f(x)\,dx = r\,\mathbb{P}\{g(X) \geq r\}. \tag{46}$$

5.3 Jensen's Inequality

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

$$\left.\begin{array}{l} x_1 \leq x_2 \\ 0 \leq t \leq 1 \end{array}\right\} \Rightarrow \phi(tx_1 + (1-t)x_2) \leq t\phi(x_1) + (1-t)\phi(x_2), \tag{47}$$

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

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X). \tag{48}$$

The conditions for convexity given in (47) can be expressed globally as an inequality relation between a function and its tangent line at an arbitrary point, x, so that we have

$$ax + b \leq \phi(x), \tag{49}$$

for the appropriate values, a and b. If we choose the particular point, $x_0 = \mathbb{E}X$, then the inequality in (49) leads to the relations,

$$\begin{aligned} \mathbb{E}\phi(X) = \int_{\mathbb{R}} \phi(x)p(x)\,dx \geq \int_{\mathbb{R}} (ax+b)p(x)\,dx = \\ a\int_{\mathbb{R}} xp(x)\,dx + b\int_{\mathbb{R}} p(x)\,dx = ax_0 + b = \phi(x_0) = \phi(\mathbb{E}X). \end{aligned} \tag{50}$$

6 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 §§6.1.1 - 6.1.3, 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.

6.1 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**,

$$X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots = \sum_{n=1}^{\infty} \frac{X^n}{n!} \equiv e^X, \tag{51}$$

which is well-defined provided the moments exist.

6.1.1 Moment-Generating Functions

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

$$M_X(t) \equiv \mathbb{E}e^{tX} = \sum_{n=1}^{\infty} \frac{\mathbb{E}X^n}{n!} t^n, \tag{52}$$

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

$$\mathbb{E}X^n = \frac{d^n}{dt^n} M_X(t)\bigg|_{t=0} = M_X^{(n)}(0), \tag{53}$$

for which the n^{th} term in the series is the n^{th} moment of the distribution. Notice that the expression in (52) 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,

$$M_{cX}(t) = \mathbb{E}e^{ctX} = M_X(ct) \tag{54}$$

while the joint moment-generating function for independent random variables, X and Y, (*cf* §4), factors into the product of the marginal moment-generating functions,

$$M_{X,Y}(s,t) = \mathbb{E}e^{sX+tY} = \mathbb{E}e^{sX}e^{tY} = \mathbb{E}e^{sX}\mathbb{E}e^{tY} = M_X(s)M_Y(t). \tag{55}$$

The first equality in the chain in (55) 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.

6.1.2 Characteristic Functions

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

$$\phi_X(t) \equiv \mathbb{E}e^{itX}. \tag{56}$$

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,

$$\phi_{cX}(t) = \mathbb{E}e^{ictX} = \phi_X(ct) \tag{57}$$

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

$$\phi_{X,Y}(s,t) = \mathbb{E}e^{i(sX+tY)} = \mathbb{E}e^{isX}e^{itY} = \mathbb{E}e^{isX}\mathbb{E}e^{itY} = \phi_X(s)\phi_Y(t). \tag{58}$$

6.1.3 Cumulants

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

$$K_X(t) = \ln M_X(t) \tag{59}$$

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,

$$\begin{aligned} K_X(t) &= \left(t\mathbb{E}X + \frac{t^2}{2}\mathbb{E}X^2 + \cdots\right) + \frac{1}{2}\left(t\mathbb{E}X + \cdots\right)^2 + \cdots \\ &= t\mathbb{E}X + \frac{t^2}{2}\left(\left(\mathbb{E}X\right)^2 - \mathbb{E}X^2\right) + \cdots . \end{aligned} \tag{60}$$

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,

$$K_{cX}(t) = K_X(ct), \tag{61}$$

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

$$K_{X,Y}(s,t) = K_X(s) + K_Y(t). \tag{62}$$

6.1.4 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,

$$\left.\begin{array}{l} X \to \mathbf{X} \\ t \to \mathbf{t} \end{array}\right\} \Rightarrow \begin{cases} M_{\mathbf{X}}(\mathbf{t}) \equiv \mathbb{E}e^{\mathbf{t}^\top \mathbf{X}} \\ K_{\mathbf{X}}(\mathbf{t}) \equiv \ln M_{\mathbf{X}}(\mathbf{t}) \\ \phi_{\mathbf{X}}(\mathbf{t}) \equiv \mathbb{E}e^{i\mathbf{t}^\top \mathbf{X}} \end{cases} \tag{63}$$

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

6.2 Transformations

6.2.1 General Transformation

$$Y = g(X) \tag{64}$$

$$\text{increasing function, } g: F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(g(X) \le Y = \mathbb{P}(X \le g^{-1}(y)) = F_X(g^{-1}(y)) \tag{65}$$

$$\text{decreasing function, } g: F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(g(X) \le Y = \mathbb{P}(X \ge g^{-1}(y)) = 1 - F_X(g^{-1}(y)) \tag{66}$$

$$p_Y(y) = \frac{d}{dy}F_Y(y) = p_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right| \tag{67}$$

6.2.2 Scale-location Adjustment

$$X \sim p(x) \Rightarrow \alpha + \beta X \sim \frac{1}{\beta}p(\alpha + \beta x) \tag{68}$$

6.2.3 Sum of Random Variables

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

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) \Rightarrow p_{X+Y}(z) = \int_{-\infty}^{\infty} p_X(x)p_Y(z-x)\,dx \tag{69}$$

7 Common Functions

Many of the common distributions used as continuous probabilistic models, described in detail below in §8.2, can be expressed in closed form using functions defined through definite integrals.

7.1 The Error Function

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

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\,dt \tag{70}$$

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) \tag{71}$$

which are used for closed-form expression of cumulative Gaussian probabilities, as shown in §8.2.1. Notice that the argument, x, appears as the limit of integration.

7.2 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

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt, \tag{72}$$

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,

$$\begin{aligned}
\Gamma(x) &= \int_0^{\infty} t^{x-1} e^{-t}\,dt \\
&= -t^{x-1}e^{-t}\big|_0^{\infty} - \int_0^{\infty} (x-1)t^{x-2}(-e^{-t})\,dt \\
&= (x-1)\int_0^{\infty} t^{x-2}e^{-t}\,dt \\
&= (x-1)\Gamma(x-1).
\end{aligned} \tag{73}$$

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

$$\Gamma(n) = (n-1)! \tag{74}$$

7.3 The Beta Function

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

$$\mathrm{B}(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt \tag{75}$$

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

$$\begin{aligned}\Gamma(x)\Gamma(y) &= \int_0^\infty s^{x-1}e^{-s}\,ds \int_0^\infty t^{y-1}e^{-t}\,dt \\ &= \int_0^\infty \int_0^\infty s^{x-1}t^{y-1}e^{-(s+t)}\,ds\,dt && \begin{cases} s = uv \\ t = u(1-v) \end{cases} \\ &= \int_{u=0}^\infty \int_{v=0}^1 (uv)^{x-1}(u(1-v))^{y-1}e^{-u}u\,du\,dv && |J| = u \\ &= \int_0^\infty e^{-u}u^{x+y-1}\,du \int_0^1 v^{x-1}(1-v)^{y-1}\,dv \\ &= \Gamma(x+y)\,\mathrm{B}(x,y) \end{aligned} \tag{76}$$

which can be rearrange to yield,

$$\mathrm{B}(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}. \tag{77}$$

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

$$\begin{aligned}\mathrm{B}\left(\frac{1}{2},\frac{1}{2}\right) &= \int_0^1 t^{-\frac{1}{2}}(1-t)^{-\frac{1}{2}}\,dt \\ &= 2\int_0^{\frac{\pi}{2}} \frac{\cos\theta\sin\theta}{\cos\theta\sin\theta}\,d\theta \\ &= \pi \end{aligned} \tag{78}$$

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

$$\mathrm{B}\left(\frac{1}{2},\frac{1}{2}\right) = \frac{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(1)} \Rightarrow \Gamma\left(\frac{1}{2}\right) = \sqrt{\mathrm{B}\left(\frac{1}{2},\frac{1}{2}\right)} = \sqrt{\pi}. \tag{79}$$

7.3.1 The Multivariate Beta Function

The relation between beta and gamma functions in (77) can be generalized for multivariate argument,

$$\mathrm{B}(\alpha_1,\ldots,\alpha_n) = \frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^n \alpha_i\right)}. \tag{80}$$

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

$$\mathrm{B}(\alpha_1,\ldots,\alpha_n) = \frac{\Gamma(\alpha_j)\prod_{i\neq j}\Gamma(\alpha_i)}{\Gamma\left(\alpha_j + \sum_{i\neq j}\alpha_i\right)} = \frac{\prod_{i\neq j}\Gamma(\alpha_i)}{\Gamma(\sum_{i\neq j}\alpha_i)}\,\mathrm{B}\left(\alpha_j, \sum_{i\neq j}\alpha_i\right). \tag{81}$$

8 Common Distributions

8.1 Discrete Distributions

8.1.1 Sampling With Replacement

8.1.1.1 Bernoulli

$$\mathrm{Ber}(p) \equiv f(k|p) = p^k(1-p)^{1-k}, k \in \{0,1\} \tag{82}$$

8.1.1.2 Binomial

$$X_i \sim \mathrm{Ber}(p) \Rightarrow Y = \sum_{i=1}^{n} X_i \sim \mathrm{Bin}(n, p) \tag{83}$$

$$\mathrm{Bin}(n, p) \equiv f(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}, k \in \{0, \cdots, n\} \tag{84}$$

8.1.1.3 Negative Binomial

$$\mathrm{NB}(k|r, p) = \mathrm{Bin}(k|k+r-1, p)\,\mathrm{Ber}(0|p) \tag{85}$$

$$\mathrm{NB}(k|r, p) \equiv f(k|r, p) = \binom{k+r-1}{k} p^k (1-p)^{r-k}, k \in \mathbb{N} \tag{86}$$

8.1.1.4 Geometric

$$\mathrm{Geo}(k|p) = \mathrm{NB}(k|1, 1-p) = p(1-p)^{k-1}, k \in \mathbb{N} \tag{87}$$

8.1.1.5 Poisson

$$\mathrm{Poi}(\lambda) = f(k|\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}, k \in \mathbb{N} \tag{88}$$

$$\begin{aligned}
\mathrm{Bin}\left(n, \frac{\lambda}{n}\right) &= f\left(k \,\middle|\, n, \frac{\lambda}{n}\right) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{k}\right)^{n-k} \\
&= \frac{n \cdot n-1 \cdot \cdots \cdot n-k+1}{k!} \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^k \\
&= \left[\left(\frac{n}{n}\right) \cdot \left(\frac{n-1}{n}\right) \cdot \cdots \cdot \left(\frac{n-k+1}{n}\right)\right] \cdot \left[\left(1 - \frac{\lambda}{n}\right)^{-k}\right] \cdot \left[\frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n\right]
\end{aligned} \tag{89}$$

$$\lim_{n \to \infty} \mathrm{Bin}\left(n, \frac{\lambda}{n}\right) = \frac{\lambda^k}{k!} e^{-\lambda} \equiv \mathrm{Poi}(\lambda) \tag{90}$$

8.1.1.6 Multinomial

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

$$\mathrm{Mul}(n, p_1, \cdots, p_k) \equiv f(x_1, \cdots, x_k | n, p_1, \cdots, p_k) = \frac{n}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i} = \frac{\Gamma(1 + \sum_{i=1}^{k} x_i)}{\prod_{i=1}^{k} \Gamma(1 + x_i)} \prod_{i=1}^{k} p_i^{x_i} \tag{91}$$

$$X_1, \cdots, X_k \sim \mathrm{Mul}(n, p_1, \cdots, p_k) \Rightarrow \begin{cases} \mathbb{E}X_i = np_i \\ \mathbb{V}X_i = np_i(1-p_i) \\ \mathbb{C}(X_i, X_j) = -np_ip_j, i \neq j \end{cases} \tag{92}$$

8.1.2 Sampling Without Replacement

8.1.2.1 Hypergeometric

$$\text{Hyp}(n, N, K) \equiv f(k|n, N, K) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{93}$$

8.1.2.2 Multivariate Hypergeometric

$$\left.\begin{array}{l}\mathbf{k} = (k_1, \ldots, k_n)^\top \\ \mathbf{K} = (K_1, \ldots, K_n)^\top\end{array}\right\} \Rightarrow \text{MHG}(\mathbf{k}|\mathbf{K}) = \frac{\binom{K_1}{k_1}\cdots\binom{K_n}{k_n}}{\binom{\sum_{i=1}^n K_i}{\sum_{i=1}^n k_i}} \tag{94}$$

8.2 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.

8.2.1 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 §10.2.2. 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.

8.2.1.1 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

$$\text{probability density:} \quad \text{N}(\mu, \sigma^2) \equiv p_\text{N}(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{95}$$

$$\text{cumulative distribution:} \quad F_\text{N}(x|\mu, \sigma^2) = \int_{-\infty}^{x} p_\text{N}(x|\mu, \sigma^2)\, dx = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right) \tag{96}$$

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

$$g(\alpha) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\alpha\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \frac{1}{\sqrt{\alpha}}, \tag{97}$$

and the central moments are derived by direct calcuation,

$$\mathbb{E}(X-\mu)^{2n-1} = 0 \tag{98}$$

$$\mathbb{E}(X-\mu)^{2n} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^{2n} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = (-2\sigma^2)^n \left.\frac{d^n}{d\alpha^n} g(\alpha)\right|_{\alpha=1} = (2n-1)!!\,\sigma^{2n} \tag{99}$$

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.

$$\begin{aligned}
M_{\mathrm{N}(\mu,\sigma^2)}(t) \equiv \mathbb{E}e^{tX} &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2-tx}{2\sigma^2}}\,dx \\
&= e^{t\mu+\frac{1}{2}t^2\sigma^2}\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{(x-(\mu+t\sigma^2))^2}{2\sigma^2}}\,dx \\
&= e^{t\mu+\frac{1}{2}t^2\sigma^2}
\end{aligned} \tag{100}$$

$$\phi_{\mathrm{N}(\mu,\sigma^2)}(t) \equiv \mathbb{E}e^{itX} = e^{it\mu-\frac{1}{2}t^2\sigma^2} \tag{101}$$

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

$$\left.\begin{aligned} X \sim \mathrm{N}(\mu,\sigma^2) \Rightarrow M_X = e^{t\mu+\frac{1}{2}t^2\sigma^2} \\ Y \sim \mathrm{N}(\nu,\tau^2) \Rightarrow M_Y = e^{t\nu+\frac{1}{2}t^2\tau^2} \end{aligned}\right\} \Rightarrow M_{X+Y} = e^{t(\mu+\nu)+\frac{1}{2}t^2(\sigma^2+\tau^2)} \Rightarrow X+Y \sim \mathrm{N}(\mu+\nu,\sigma^2+\tau^2). \tag{102}$$

8.2.1.2 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

$$Z \sim \mathrm{N}(0,1). \tag{103}$$

All other derived quantities and functions are calculated similarly.

8.2.1.3 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, $\boldsymbol{\Sigma}$,

$$\mathbf{X} = \begin{pmatrix} X_1 \sim \mathrm{N}(\mu_1,\sigma_1^2) \\ \vdots \\ X_n \sim \mathrm{N}(\mu_n,\sigma_n^2) \end{pmatrix} \Rightarrow \begin{cases} \boldsymbol{\mu} = \begin{pmatrix} \mathbb{E}X_1 \\ \vdots \\ \mathbb{E}X_n \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \\ \\ \boldsymbol{\Sigma} = \begin{pmatrix} \mathbb{V}X_1 & \cdots & \mathbb{C}(X_1,X_n) \\ \vdots & \ddots & \vdots \\ \mathbb{C}(X_1,X_n) & \cdots & \mathbb{V}X_n \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{x_1x_n} \\ \vdots & \ddots & \vdots \\ \sigma_{x_1x_n} & \cdots & \sigma_n^2 \end{pmatrix} \end{cases} \tag{104}$$

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 (104), the multivariate Gaussian distribution takes the form,

$$\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \det\boldsymbol{\Sigma}}}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{105}$$

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 all interactio and indentity covariance matrix,

$$\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\mathbf{z}^\top \mathbf{z}} \tag{106}$$

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

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{Z}, \tag{107}$$

frow which we derive

$$\left.\begin{aligned} &\mathbb{E}\mathbf{X} \equiv \mathbb{E}(\boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{Z}) = \boldsymbol{\mu} \\ &\mathbb{V}\mathbf{X} \equiv \mathbb{E}\left(\mathbf{X} - \mathbb{E}\mathbf{X}\right)\left(\mathbf{X} - \mathbb{E}\mathbf{X}\right)^\top = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top \end{aligned}\right\} \Rightarrow \mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top) \tag{108}$$

Similarly we can recover the standard normal by inverting the relationship in (107),

$$\mathbf{Z} = \boldsymbol{\Gamma}^{-1}(\mathbf{X} - \boldsymbol{\mu}), \tag{109}$$

so that

$$\left.\begin{aligned} &\mathbb{E}\mathbf{Z} \equiv \mathbb{E}\boldsymbol{\Gamma}(\mathbf{X} - \boldsymbol{\mu}) = \mathbf{0} \\ &\mathbb{V}\mathbf{Z} \equiv \mathbb{E}\left(\mathbf{Z} - \mathbb{E}\mathbf{Z}\right)\left(\mathbf{Z} - \mathbb{E}\mathbf{Z}\right)^\top = \mathbb{E}\mathbf{Z}\mathbf{Z}^\top = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-\top} = \mathbf{I} \end{aligned}\right\} \Rightarrow \mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}) \tag{110}$$

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,

$$\boldsymbol{\Sigma} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top \Rightarrow \boldsymbol{\Sigma}^{-1} = \mathbf{Q}\mathbf{D}^{-1}\mathbf{Q}^\top. \tag{111}$$

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

$$\boldsymbol{\Sigma} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top \Rightarrow \mathbf{D} = \mathrm{diag}(\zeta_1^2, \ldots, \zeta_n^2) \tag{112}$$

$$\boldsymbol{\Sigma}^{-1} = \mathbf{Q}\mathbf{D}^{-1}\mathbf{Q}^\top \Rightarrow \mathbf{D}^{-1} = \mathrm{diag}\left(\frac{1}{\zeta_1^2}, \ldots, \frac{1}{\zeta_n^2}\right). \tag{113}$$

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 (107) by further decomposing the diagonal matrix into a square root:

$$\boldsymbol{\Sigma} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top = \mathbf{Q}\mathbf{D}^{\frac{1}{2}}\mathbf{D}^{\frac{\top}{2}}\mathbf{Q}^\top = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top \Rightarrow \boldsymbol{\Gamma} = \mathbf{Q}\mathbf{D}^{\frac{1}{2}}, \qquad \mathbf{D}^{\frac{1}{2}} = \mathrm{diag}(|\zeta_1|, \ldots, |\zeta_n|). \tag{114}$$

If we now introduce a change of basis (*cf LAN* §3.9.1) 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

$$\mathbf{y} = \mathbf{Q}^\top(\mathbf{x} - \boldsymbol{\mu}), \tag{115}$$

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,

$$\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \frac{1}{2}\mathbf{y}^\top \mathbf{D}^{-1}\mathbf{y} = \sum_{i=1}^{n} \frac{y_i^2}{2\zeta_i^2} \tag{116}$$

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

$$\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} = \prod_{i=1}^{n} \frac{1}{\zeta_i \sqrt{2\pi}} e^{-\frac{y_i^2}{2\zeta_i^2}} = \prod_{i=1}^{n} f_{\mathrm{N}}(y_i|0, \zeta_i^2). \tag{117}$$

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

8.2.1.4 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

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \Rightarrow \begin{cases} \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \\ \\ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \end{cases} \tag{118}$$

Following the discussion in §4 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,

$$(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) = \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu} \\ \mathbf{x}_2 - \boldsymbol{\mu} \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu} \\ \mathbf{x}_2 - \boldsymbol{\mu} \end{pmatrix} \tag{119}$$

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

$$\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} I_p & 0 \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & I_q \end{pmatrix} \begin{pmatrix} \left(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)^{-1} & 0 \\ 0 & \boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix} \begin{pmatrix} I_p & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ 0 & I_q \end{pmatrix} \tag{120}$$

which can be expanded to yield

$$\begin{aligned} &(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ &= \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu} \\ \mathbf{x}_2 - \boldsymbol{\mu} \end{pmatrix}^\top \begin{pmatrix} I_p \\ \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \end{pmatrix} \left(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)^{-1} \begin{pmatrix} I_p \\ -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix}^\top \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu} \\ \mathbf{x}_2 - \boldsymbol{\mu} \end{pmatrix} + (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \end{aligned} \tag{121}$$

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,

$$\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathrm{N}(\boldsymbol{\mu}_1 - (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})\ \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}). \tag{122}$$

This yields the desired result

$$\text{Marginal Distribution: } \mathbf{X}_2 \sim \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \tag{123}$$

$$\text{Conditional Distribution: } \mathbf{X}_1 | (\mathbf{X}_2 = \mathbf{x}_2) \sim \mathrm{N}(\boldsymbol{\mu}_1 - (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}) \tag{124}$$

8.2.1.5 Mean and Variance of IID Normal Random Variables

$$\left.\begin{array}{l}\mathbf{X} = (X_1, \cdots, X_n)^\top, X_i \sim \mathrm{N}(\mu, \sigma^2) \\ \mathbf{1}_n = (1, \cdots, 1)^\top\end{array}\right\} \Rightarrow \mathbf{X} \sim \mathrm{N}(\mu \mathbf{1}_n, \sigma^2 I_n) \tag{125}$$

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

$$\mathbf{Y} = (\mathbf{Y}_1, \cdots, \mathbf{Y}_n)^\top = Q^\top \mathbf{X} \Rightarrow \mathbf{Y} \sim \mathrm{N}(\mu \mathbf{1}_n, \sigma^2 Q^\top I_n Q) = \mathrm{N}(\mu \mathbf{1}_n, \sigma^2 I_n) \tag{126}$$

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

$$\hat{\mu}\mathbf{1}_n = \mathbf{1}_n \frac{1}{n} \mathbf{1}_n^\top \mathbf{x} = \mathbf{1}_n (\mathbf{1}_n^\top \mathbf{1}_n)^{-1} \mathbf{1}_n^\top \mathbf{x} = P_{\mathbf{1}_n} \mathbf{x} \tag{127}$$

$$\hat{\sigma}^2 = \frac{1}{n-1} (\mathbf{x} - \hat{\mu}\mathbf{1}_n)^\top (\mathbf{x} - \hat{\mu}\mathbf{1}_n) = \frac{1}{n-1} (\mathbf{x} - P_{\mathbf{1}_n}\mathbf{x})^\top (\mathbf{x} - P_{\mathbf{1}_n}\mathbf{x}) = \frac{1}{n-1} \mathbf{x}^\top (I_n - P_{\mathbf{1}_n}) \mathbf{x} \tag{128}$$

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

Sums of Gaussian are Gaussian

$$\sqrt{n}\hat{\mu} \sim \mathrm{N}(\mu, \sigma^2) \tag{129}$$

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

$$(n-1)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2} \mathbf{x}^\top (I_n - P_{\mathbf{1}_n}) \mathbf{x} = \sum_{i=2}^{n} \frac{y_i^2}{\sigma^2} = \sum_{i=2}^{n} z_i^2 \sim \chi^2_{n-1} \tag{130}$$

8.2.2 Gamma-Derived Distributions

8.2.2.1 Gamma

$$\Gamma(\alpha, \beta) \equiv f_\Gamma(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},\ x \geq 0 \tag{131}$$

$$\begin{aligned}\Gamma(\alpha) &= \int_0^\infty t^{\alpha-1} e^{-t}\, dt \\ &= \int_0^\infty (\beta x)^{\alpha-1} e^{-\beta x} \beta\, dx \qquad t = \beta x \\ &= \int_0^\infty \beta^\alpha x^{\alpha-1} e^{-\beta x}\, dx\end{aligned} \tag{132}$$

$$
\begin{aligned}
\mathbb{E}X^n &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^k x^{\alpha-1} e^{-\beta x}\,dx \\
&= \frac{\beta^\alpha}{\beta^{\alpha+k}} \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \int_0^\infty \frac{\beta^{\alpha+k}}{\Gamma(\alpha+k)} x^{\alpha+k-1} e^{-\beta x}\,dx \\
&= \frac{1}{\beta^k} \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \int_0^\infty f_\Gamma(x|\alpha+k,\beta)\,dx \\
&= \frac{1}{\beta^k} \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}
\end{aligned} \tag{133}
$$

$$
\mathbb{E}X = \frac{\alpha}{\beta} \tag{134}
$$

$$
\mathbb{V}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{(\alpha+1)\alpha}{\beta^2} - \frac{\alpha^2}{\beta^2} = \frac{\alpha}{\beta^2} \tag{135}
$$

$$
\begin{aligned}
M_{\Gamma(\alpha,\beta)} \equiv \mathbb{E}e^{tX} &= \int_0^\infty e^{tx} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta-t)x}\,dx \\
&= \frac{\beta^\alpha}{(\beta-t)^\alpha} \int_0^\infty \frac{(\beta-t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta-t)x}\,dx \\
&= \left(\frac{\beta}{\beta-t}\right)^\alpha \int_0^\infty f_\Gamma(x|\alpha,\beta-t)\,dx \\
&= \left(\frac{\beta}{\beta-t}\right)^\alpha
\end{aligned} \tag{136}
$$

$$
X_i \sim \Gamma(\alpha_i,\beta) \Rightarrow \sum_{i=1}^n X_i \sim \Gamma\left(\sum_{i=1}^n \alpha_i, \beta\right) \tag{137}
$$

8.2.2.2 Chi-square

$$
X \sim \mathrm{N}(0,1) \Rightarrow X^2 \sim \chi^2 = \Gamma\left(\frac{1}{2},\frac{1}{2}\right) \tag{138}
$$

$$
\mathbb{P}\{X^2 \le x\} = \mathbb{P}\{-\sqrt{x} \le X \le \sqrt{x}\} = \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt \tag{139}
$$

$$
\begin{aligned}
\chi^2 \equiv f_{\chi^2}(x) &= \frac{d}{dx} \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt \\
&= \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x}{2}}\right)\left(\frac{1}{2}x^{-\frac{1}{2}}\right) - \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x}{2}}\right)\left(-\frac{1}{2}x^{-\frac{1}{2}}\right) \\
&= \frac{1}{\sqrt{2\pi}} x^{\frac{1}{2}-1} e^{-\frac{x}{2}} \\
&= \Gamma\left(\frac{1}{2},\frac{1}{2}\right)
\end{aligned} \tag{140}
$$

$$
X_i \sim \mathrm{N}(0,1) \Rightarrow \sum_{i=1}^n X_i^2 \sim \chi_n^2 = \Gamma\left(\frac{n}{2},\frac{1}{2}\right) \tag{141}
$$

$$X \sim \chi^2_n \Rightarrow \begin{cases} \mathbb{E}X = n \\ \mathbb{V}X = \frac{n}{2} \end{cases} \tag{142}$$

$$\boldsymbol{\Sigma} = \mathbf{QDQ}^\top = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top \Rightarrow \mathbf{Z} = \boldsymbol{\Gamma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \tag{143}$$

$$\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow (\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) = \mathbf{Z}^\top \mathbf{Z} \sim \chi^2_n \tag{144}$$

8.2.2.3 Inverse Gamma

8.2.3 Beta-Derived Distributions

8.2.3.1 Beta

$$\mathrm{B}(\alpha, \beta) \equiv f_{\mathrm{B}}(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 \leq x \leq 1 \tag{145}$$

$$\begin{aligned}
\mathbb{E}X^k &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^k x^{\alpha-1}(1-x)^{\beta-1}\,dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+k)\Gamma(\beta)}{\Gamma(\alpha+\beta+k)} \int_0^1 \frac{\Gamma(\alpha+\beta+k)}{\Gamma(\alpha+k)\Gamma(\beta)} x^{\alpha+k-1}(1-x)^{\beta-1}\,dx \\
&= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+k)} \int_0^1 f_{\mathrm{B}}(x|\alpha+k, \beta)\,dx \\
&= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+k)}
\end{aligned} \tag{146}$$

$$\mathbb{E}X = \frac{\alpha}{\alpha+\beta} \tag{147}$$

$$\mathbb{V}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2} = \frac{\alpha\beta}{(\alpha+\beta)^2} \tag{148}$$

8.2.3.2 Dirichlet

$$\mathrm{Dir}(\alpha_1, \cdots, \alpha_n) \equiv f_{\mathrm{D}}(x_1, \cdots, x_n|\alpha_1, \cdots, \alpha_n) = \frac{\prod_{i=1}^n x_i^{\alpha_i-1}}{\mathrm{B}(\alpha_1, \cdots, \alpha_n)}, \begin{cases} 0 \leq x_i \leq 1 \\ \sum_{i=1}^n x_i = 1 \end{cases} \tag{149}$$

marginal distributions

$$f_D(x_i|\alpha_1, \cdots, \alpha_n) = f_{\mathrm{B}}\left(x_i \middle| \alpha_i, \sum_{j \neq i} \alpha_j \right) = \frac{x_i^{\alpha_i-1}(1-x_i)^{\sum_{j\neq i}\alpha_j - 1}}{\mathrm{B}(\alpha_i, \sum_{j\neq i}\alpha_j)} \tag{150}$$

$$\mathbb{E}X_i = \frac{\alpha_i}{\sum_{j=1}^n \alpha_j} \tag{151}$$

$$\mathbb{V}X_i = \frac{\alpha_i \sum_{j\neq i}\alpha_j}{\left(\sum_{j=1}^n \alpha_j\right)^2} \tag{152}$$

8.2.4 Distributions of Ratios of Standard Normal Random Variables

8.2.4.1 F-Distribution

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

$$F(k,m) \equiv f_F(x|k,m) = \frac{\Gamma(\frac{k+m}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}\left(\frac{k}{m}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1}\left(1+\frac{k}{m}x\right)^{-\frac{k+m}{2}} \tag{153}$$

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

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

$$= \iint_{\{U\le\frac{k}{m}xV\}} f_{\chi_k^2}(u) f_{\chi_m^2}(v)\,du\,dv = \int_0^\infty \int_0^{\frac{k}{m}xv} f_{\chi_k^2}(u) f_{\chi_m^2}(v)\,du\,dv \tag{154}$$

$$\begin{aligned}
F(k,m) \equiv f_F(x|k,m) &= \frac{d}{dx}\int_0^\infty \int_0^{\frac{k}{m}xv} f_{\chi_k^2}(u) f_{\chi_m^2}(v)\,du\,dv = \int_0^\infty \frac{d}{dx}\left(\int_0^{\frac{k}{m}xv} f_{\chi_k^2}(u)\,du\right) f_{\chi_m^2}(v)\,dv \\
&= \frac{k}{m}\int_0^\infty f_{\chi_k^2}\left(\frac{k}{m}xv\right) f_{\chi_m^2}(v)\,v\,dv \\
&= \frac{k}{m}\int_0^\infty \left(\frac{\frac{1}{2}^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)}\left(\frac{k}{m}xv\right)^{\frac{k}{2}-1} e^{-\frac{\frac{k}{m}xv}{2}}\right)\left(\frac{\frac{1}{2}^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)} v^{\frac{m}{2}-1} e^{-\frac{v}{2}}\right) v\,dv \\
&= \frac{k^{\frac{k}{2}}}{m}\int_0^\infty \frac{\left(\frac{1}{2}\right)^{\frac{k+m}{2}} x^{\frac{k}{2}-1}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} v^{\frac{k+m}{2}-1} e^{-\frac{1}{2}v\left(\frac{k}{m}t+1\right)}\,dv \\
&= \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)}\left(\frac{k}{m}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1}\left(\frac{1}{\frac{k}{m}x+1}\right)^{\frac{k+m}{2}} \int_0^\infty \frac{\left(\frac{t+1}{2}\right)^{\frac{k+m}{2}}}{\Gamma\left(\frac{k+m}{2}\right)} v^{\frac{k+m}{2}-1} e^{-v\frac{t+1}{2}}\,dv \\
&= \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)}\left(\frac{k}{m}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1}\left(1+\frac{k}{m}x\right)^{-\frac{k+m}{2}} \int_0^\infty f_\Gamma\left(v\,\middle|\,\frac{k+m}{2},\frac{t+1}{2}\right)dv \\
&= \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)}\left(\frac{k}{m}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1}\left(1+\frac{k}{m}x\right)^{-\frac{k+m}{2}}
\end{aligned} \tag{155}$$

8.2.4.2 T-Distribution

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

$$T(m) \equiv f_T(x|m) = \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{m}{2})}\frac{1}{\sqrt{m}}\left(1+\frac{x^2}{m}\right)^{-\frac{m+1}{2}} \tag{156}$$

$$\left.\begin{matrix} U = U_i^2 \sim \chi^2 \\ V = \sum_{j=1}^{m} V_j^2 \sim \chi_m^2 \end{matrix}\right\} \Rightarrow \mathbb{P}\left\{-x \le \frac{\sqrt{U}}{\sqrt{\frac{1}{m}V}} \le x\right\} = \mathbb{P}\left\{\frac{U}{\frac{1}{m}V} \le x^2\right\} = \int_0^{x^2} f_F(v|1,m)\,dv \tag{157}$$

$$T(m) \equiv f_T(x|m) = \frac{1}{2}\frac{d}{dx}\int_0^{x^2} f_F(v|1,m)\,dv = x f_F\left(x^2|1,m\right)$$
$$= \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{m}{2}\right)}\frac{1}{\sqrt{m}}\left(1+\frac{x^2}{m}\right)^{-\frac{m+1}{2}} \tag{158}$$

8.2.4.3 Cauchy

$$\left.\begin{matrix} U \\ V \end{matrix}\right\} \sim Z = \mathrm{N}(0,1) \Rightarrow \frac{U}{V} \sim \mathrm{Cau}(0,1) \equiv f_C(x) = \frac{1}{\pi}\frac{1}{x^2+1} \tag{159}$$

$$\mathbb{P}\left\{\frac{U}{V} \le x\right\} = \mathbb{P}\left\{U \le xV\right\} = \iint_{\{u \le xv\}} f_{\mathrm{N}}(u|0,1) f_{\mathrm{N}}(v|0,1)\,du\,dv$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{xv} f_{\mathrm{N}}(u|0,1) f_{\mathrm{N}}(v|0,1)\,du\,dv \tag{160}$$

$$\begin{aligned}\mathrm{Cau}(0,1) \equiv f_{\mathrm{C}}(x) &= \frac{d}{dx}\mathbb{P}\left\{\frac{U}{V} \le x\right\} = \frac{d}{dx}\int_0^{\infty}\int_0^{xv} f_{\mathrm{N}}(u|0,1) f_{\mathrm{N}}(v|0,1)\,du\,dv \\ &= \int_{-\infty}^{\infty}\left(\frac{d}{dx}\int_{-\infty}^{xv} f_{\mathrm{N}}(u|0,1)\,du\right) f_{\mathrm{N}}(v|0,1)\,dv = \int_{-\infty}^{\infty} f_{\mathrm{N}}(xv|0,1) f_{\mathrm{N}}(v|0,1)\,dv \\ &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2v^2}{2}}\frac{1}{2\pi}e^{-\frac{v^2}{2}}v\,dv = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\frac{v^2(x^2+1)}{2}}v\,dv \\ &= \frac{1}{2\pi}\frac{1}{x^2+1}\int_{-\infty}^{\infty} e^{-t}\,dt = \frac{1}{\pi}\frac{1}{x^2+1}\int_0^{\infty} e^{-t}\,dt \\ &= \frac{1}{\pi}\frac{1}{x^2+1}\end{aligned} \tag{161}$$

$$\mathrm{Cau}(0,1) \equiv T(1) \tag{162}$$

8.2.5 Other Common Distributions

8.2.5.1 Exponential

$$\mathrm{Exp}(\lambda) \equiv f_{\mathrm{E}}(x|\lambda) = \lambda e^{-\lambda x}, \qquad x \ge 0 \tag{163}$$

memoryless:

$$\mathbb{P}\{x > s+t | x > s\} = \mathbb{P}\{x > t\} \tag{164}$$

$$h(t) = \frac{f(t)}{1 - \int_0^t f(x)\,dx} = \lambda \tag{165}$$

8.2.5.2 Pareto

$$\mathrm{Par}(\alpha, x_m) \equiv f_{\mathrm{P}}(x|\alpha, x_m) = \begin{cases}\frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}, & x \ge x_m \\ 0, & x < x_m\end{cases} \tag{166}$$

hazard rate (burn-in period)

$$h(t) = \frac{\alpha}{t} \tag{167}$$

8.2.5.3 Weibull

8.2.5.4 Uniform

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

8.2.6 Exponential Family

$$f(x|\theta) = h(x)g(\theta)e^{\eta(\theta)T(x)} \tag{169}$$

9 Order Statistics

The probability distributions of cumulative probability rank, or 'percentile', of finite samples taken with replacement from *arbitrary* distributions are known as **order statitistics**. 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, $X_i \sim \mathrm{Uni}(0,1)$. The random variables sorted in increasing order, designated as $X_{(1)}, \cdots, X_{(n)}$,

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

Treating the infinitesimal interval as finite-sized, the frequency with which each independend sample falls within each interval is governed by multinomial statistics, §8.1.1.6,

$$\frac{n!}{(k-1)!1!(n-k)!}u^{k-1} \cdot du \cdot (1-u-du)^{n-k} \approx \frac{n!}{(k-1)!(n-k)!}u^{k-1} \cdot (1-u)^{n-k} \cdot du \tag{170}$$

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

$$X_{(k)} \sim \mathrm{B}(k, n+1-k). \tag{171}$$

10 Asymptotic Limits

10.1 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:

$$X : \Omega \to \mathbb{R} \tag{172}$$

$$X^{-1} : \mathcal{B} \to \mathcal{F} \tag{173}$$

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.

10.1.1 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

$$X^{-1}[-\infty, x) = \{\omega \in \Omega : X(\omega) < x\} \equiv \{X < x\} \tag{174}$$

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:

$$\lim_{n\to\infty} \mathbb{P}\{X_n < x\} = \mathbb{P}\{X < x\} \Leftrightarrow X_n \xrightarrow{d} X \tag{175}$$

Notice that the limit is applied to real values, and that the relation in (175) must hold for all x.

10.1.2 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, $\epsilon > 0$,

$$\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\} \equiv \{|X_n - X| > \epsilon\} \tag{176}$$

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 $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\{|X_n - X| > \epsilon\} = 0 \Leftrightarrow X_n \xrightarrow{p} X. \tag{177}$$

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

10.1.3 Convergence Almost Surely (Strong Convergence)

Similar to convergence in probability, convergence almost surely starts with the sequence of difference sets defined in (176). 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

$$\limsup_{n\to\infty}\{|X_n - X| > \epsilon\} = \bigcap_{n\in\mathbb{N}} \bigcup_{m\geq n} \{\omega \in \Omega : |X_m(\omega) - X(\omega)| > \epsilon\}. \tag{178}$$

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 $\epsilon > 0$,

$$\mathbb{P}\limsup_{n\to\infty}\{|X_n - X| > \epsilon\} = 0 \Leftrightarrow X_n \xrightarrow{as} X. \tag{179}$$

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!

10.1.4 Functions of Convergent Random Variables

The **Continuous Mapping Theorem**:

$$\left.\begin{array}{l} g \text{ is a } \textit{continuous} \text{ function} \\ X_n \xrightarrow{d} X \end{array}\right\} \Rightarrow g(X_n) \xrightarrow{d} g(X). \tag{180}$$

10.1.5 Product of Convergent Random Variables

Slutsky's Theorem:

$$\left.\begin{aligned} X_n &\xrightarrow{d} X \\ Y_n &\xrightarrow{p} c \end{aligned}\right\} \Rightarrow X_n Y_n \xrightarrow{d} cX \tag{181}$$

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

10.2 Asymptotic Limits of IID Samples

10.2.1 Law of Large Numbers

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

$$\mathbb{E}X = \mu; \tag{182}$$

$$\mathbb{V}X = \sigma^2; \tag{183}$$

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

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \Rightarrow \begin{cases} \mathbb{E}\bar{X}_n = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i = \mu; \\ \mathbb{V}\bar{X}_n = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}X_i = \frac{\sigma^2}{n}. \end{cases} \tag{184}$$

due to the linearity of the expectation operator. Furthermore, the asymptotic distribution of sample mean can be proved to converge in both weak and strong senses to the mean,

$$\textbf{Weak Law of Large Numbers} \qquad \bar{X}_n \xrightarrow{p} \mu; \tag{185}$$

$$\textbf{Strong Law of Large Numbers} \qquad \bar{X}_n \xrightarrow{as} \mu. \tag{186}$$

The weak law can be shown as a simple consequence of Chebyshev's Inequality (see §5.2). Set $g(X) = |\bar{X}_n - \mu|$, then

$$\mathbb{P}\{|\bar{X}_n - \mu| \geq \epsilon\} = \mathbb{P}\{(\bar{X}_n - \mu)^2 \geq \epsilon^2\} \leq \frac{\sigma^2}{n\epsilon^2} \Rightarrow \lim_{n\to\infty} \mathbb{P}\{|\bar{X}_n - \mu| \geq \epsilon\} = 0, \tag{187}$$

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.

10.2.2 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 asymptotic estimates on the *rate* of convergence.

10.2.2.1 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,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathrm{N}(0,1). \tag{188}$$

The characteristic equation, as described above in §6.1.2, provides a ready proof of the theorem. Given the definition of the sample mean, we construct the sum of scaled variables,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \Rightarrow \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma\sqrt{n}} \Rightarrow \begin{cases} \mathbb{E}\left(\frac{X_i-\mu}{\sigma\sqrt{n}}\right) = 0 \\ \mathbb{E}\left(\frac{X_i-\mu}{\sigma\sqrt{n}}\right)^2 = \frac{1}{n}, \end{cases} \tag{189}$$

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

$$\phi_X(t) = 1 + it\mathbb{E}X - \frac{t^2}{2}\mathbb{E}X^2 + \cdots \Rightarrow \phi_{\frac{X_i-\mu}{\sigma\sqrt{n}}}(t) = 1 - \frac{t^2}{2n} + \mathcal{O}\left(n^{-\frac{3}{2}}\right) \tag{190}$$

and the full collection to leading order is given by,

$$\phi_{\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma\sqrt{n}}}(t) = \prod_{i=1}^{n}\phi_{\frac{X_i-\mu}{\sigma\sqrt{n}}}(t) = \prod_{i=1}^{n}\left(1 - \frac{t^2}{2n} + \mathcal{O}\left(n^{-\frac{3}{2}}\right)\right). \tag{191}$$

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

$$\lim_{n\to\infty}\prod_{i=1}^{n}\left(1 - \frac{t^2}{n} + \mathcal{O}\left(n^{-\frac{3}{2}}\right)\right) = \lim_{n\to\infty}\left(1 - \frac{t^2}{n} + \mathcal{O}\left(n^{-\frac{3}{2}}\right)\right)^n = e^{-\frac{t^2}{2}} = \phi_{\mathrm{N}(0,1)}(t). \tag{192}$$

10.2.2.2 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, $\mathbf{X}$,

$$\mathbb{E}\mathbf{X} = \boldsymbol{\mu}; \tag{193}$$

$$\mathbb{V}\mathbf{X} = \mathbb{E}\mathbf{X}\mathbf{X}^\top - \mathbb{E}\mathbf{X}\,\mathbb{E}\mathbf{X}^\top = \boldsymbol{\Sigma}; \tag{194}$$

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

$$\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \Rightarrow \sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{195}$$

whose covariance matrix matches the underlying distribution.

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

$$\left.\begin{array}{l} \mathbb{E}g(\mathbf{X}) = g(\boldsymbol{\mu}) \\ \mathbb{V}g(\mathbf{X}) \approx \nabla g(\boldsymbol{\mu})^\top\boldsymbol{\Sigma}\nabla g(\boldsymbol{\mu}) \end{array}\right\} \Rightarrow \sqrt{n}\left(g(\mathbf{X}_n) - g(\boldsymbol{\mu})\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \nabla g(\boldsymbol{\mu})^\top\boldsymbol{\Sigma}\nabla g(\boldsymbol{\mu})\right). \tag{196}$$

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

$$\left.\begin{array}{l} \mathbb{E}g(\mathbf{X}) = \Gamma\boldsymbol{\mu} \\ \mathbb{V}g(\mathbf{X}) = \Gamma^\top\boldsymbol{\Sigma}\Gamma \end{array}\right\} \Rightarrow \sqrt{n}\Gamma\left(\mathbf{X}_n - \boldsymbol{\mu})\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \Gamma^\top\boldsymbol{\Sigma}\Gamma)\right). \tag{197}$$

11 Likelihood Function and Information Measures

11.1 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.

11.2 Shannon Information and Entropy

The link between physical entropy and probability theory for the **microcanonical** assembly described in §11.1 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**,

$$\mathbb{I}_S X = -\ln \mathbb{P}_X. \tag{198}$$

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,

$$\mathbb{H}X \equiv \mathbb{E}\mathbb{I}_S X = \int_{-\infty}^{\infty} p_X(x) \ln p_X \, dx. \tag{199}$$

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.

11.3 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.

11.3.1 Relative Entropy (Kullback-Leibler Divergence)

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

$$D_{\mathrm{KL}}(X||Y) \equiv \mathbb{H}X - \mathbb{H}_X Y = -\int_{-\infty}^{\infty} p_X(x) \ln \frac{p_X(x)}{p_Y(x)} \, dx. \tag{200}$$

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:

$$D_{\mathrm{KL}}(X||Y) = -\int_{-\infty}^{\infty} p_X(x) \ln \frac{p_X(x)}{p_Y(x)}\, dx \geq -\ln\left(\int_{-\infty}^{\infty} p_X(x) \frac{p_X(x)}{p_Y(x)}\, dx\right) = -\ln \int_{-\infty}^{\infty} p_Y(x)\, dx = 0. \tag{201}$$

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.

11.3.2 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,

$$\mathbb{H}(X|Y) = \mathbb{H}(X,Y) - \mathbb{H}Y. \tag{202}$$

11.3.3 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,

$$\mathbb{I}_M(X,Y) = \mathbb{H}X + \mathbb{H}Y - \mathbb{H}(X,Y) = D_{\mathrm{KL}}\left((X,Y)||\,XY\right) = \mathbb{E}_Y D_{\mathrm{KL}}\left(X|Y||\,X\right). \tag{203}$$

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.

11.4 Likelihood Function and Fisher Information

In the definitions of Shannon information and entropy, in (198) and (199), 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 $f(\mathbf{x}|\theta)$ be the joint distribution of the sample, $\mathbf{X} = (X_1, \cdots, X_n)$, for which each coordinate is taken from a distribution parametrized by θ. The **likelihood function**, $L(\theta|\mathbf{x})$, given that $\mathbf{X} = \mathbf{x}$ is observed, is defined as

$$L(\theta|\mathbf{x}) \equiv f(\mathbf{x}|\theta) \tag{204}$$

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

$$\ln L(\theta|\mathbf{x}) = \ln f(\mathbf{x}|\theta) = \ln \mathbb{P}\mathbf{X}_\theta = -\mathbb{I}_S\mathbf{X}_\theta. \tag{205}$$

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

$$L(\theta|\mathbf{x}) \equiv f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \Rightarrow \ln L(\theta|\mathbf{x}) = \sum_{i=1}^{n} \ln f(x_i|\theta). \tag{206}$$

11.4.1 Fisher Information

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

$$S(\theta|\mathbf{x}) = \frac{\partial}{\partial\theta}\ln L(\theta|\mathbf{x}) = \frac{\partial}{\partial\theta}\ln f(\mathbf{x}|\theta) = \frac{1}{f(\mathbf{x}|\theta)}\frac{\partial}{\partial\theta}f(\mathbf{x}|\theta), \tag{207}$$

which is the derivative with respect to the parameter of the Shannon information in the joint distribution. Given the two integration identities,

$$\int_D f\frac{\partial}{\partial\theta}\ln f\,d\mathbf{x} = \int_D \frac{\partial}{\partial\theta}f\,d\mathbf{x} = \frac{\partial}{\partial\theta}\int_D f\,d\mathbf{x} = 0 \tag{208}$$

$$\int_D f\left(\frac{\partial}{\partial\theta}\ln f\right)^2 d\mathbf{x} = \int_D \frac{1}{f}\left(\frac{\partial f}{\partial\theta}\right)^2 d\mathbf{x} = \int_D \left(\frac{\partial^2 f}{\partial\theta^2} - f\frac{\partial^2}{\partial\theta^2}\ln f\right) d\mathbf{x} = -\int_D f\frac{\partial^2}{\partial\theta^2}\ln f\,d\mathbf{x} \tag{209}$$

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

$$\mathbb{E}_\theta S(\theta|\mathbf{X}) = 0 \tag{210}$$

$$\mathbb{V}_\theta S(\theta|\mathbf{X}) = \mathbb{E}_\theta S(\theta|\mathbf{X})^2 - (\mathbb{E}_\theta S(\theta|\mathbf{X}))^2 = \mathbb{E}_\theta S(\theta|\mathbf{X})^2 = -\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}S(\theta|\mathbf{X})\right) \tag{211}$$

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

$$\mathbb{I}_F\mathbf{X}_\theta = \mathbb{V}S(\theta|\mathbf{X}) = -\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}S(\theta|\mathbf{X})\right) = -\mathbb{E}_\theta\frac{\partial^2}{\partial\theta^2}\ln f(\mathbf{X}|\theta). \tag{212}$$

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

$$\ln L(\theta+\Delta\theta|\mathbf{x}) \approx \ln L(\theta|\mathbf{x}) + \frac{\partial}{\partial\theta}\ln L(\theta|\mathbf{x})\Delta\theta + \frac{1}{2}\frac{\partial^2}{\partial\theta^2}\ln L(\theta|\mathbf{x})\,\Delta\theta^2, \tag{213}$$

and applying the expectation operator to the result,

$$
\begin{aligned}
-\mathbb{E}_\theta \ln L(\theta+\Delta\theta|\mathbf{X}) &\approx -\mathbb{E}_\theta \left(\ln L(\theta|\mathbf{X}) + \frac{\partial}{\partial\theta}\ln L(\theta|\mathbf{X})\Delta\theta + \frac{1}{2}\frac{\partial^2}{\partial\theta^2}\ln L(\theta|\mathbf{X})\Delta\theta^2\right) \\
&= \mathbb{E}_\theta \mathbb{I}_S\mathbf{X}_\theta - \mathbb{E}_\theta S(\theta|\mathbf{X})\Delta\theta + \frac{1}{2}\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}S(\theta|\mathbf{X})\right)\Delta\theta^2 \\
\mathbb{H}\mathbf{X}_{\theta+\Delta\theta} &\approx \mathbb{H}\mathbf{X}_\theta + \frac{1}{2}\mathbb{I}_F\mathbf{X}_\theta\Delta\theta^2
\end{aligned}
\tag{214}
$$

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,

$$
\begin{aligned}
D_{\mathrm{KL}}\left(L(\theta|\mathbf{x})||L(\theta+\Delta\theta|\mathbf{x})\right) &= \int_\Theta L(\theta|\mathbf{x})\ln\frac{L(\theta|\mathbf{x})}{L(\theta+\Delta\theta|\mathbf{x})}\,d\theta \\
&= \int_\Theta L(\theta|\mathbf{x})\left(\ln L(\theta|\mathbf{x}) - \ln L(\theta+\Delta\theta|\mathbf{x})\right)\,d\theta \\
&\approx \int_\Theta L(\theta|\mathbf{x})\left(-\frac{\partial}{\partial\theta}\ln L(\theta|\mathbf{x})\Delta\theta - \frac{1}{2}\frac{\partial^2}{\partial\theta^2}\ln L(\theta|\mathbf{x})\Delta\theta^2\right)\,d\theta \\
&= -\mathbb{E}_\theta S(\theta|\mathbf{X})\,\Delta\theta - \frac{1}{2}\mathbb{E}_\theta\frac{\partial}{\partial\theta}S(\theta|\mathbf{X})\,\Delta\theta^2 \\
&= \frac{1}{2}\mathbb{I}_F\mathbf{X}_\theta\,\Delta\theta^2
\end{aligned}
\tag{215}
$$

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

$$
\mathbb{H}\mathbf{X}_{\theta+\Delta\theta} \approx \mathbb{H}\mathbf{X}_\theta + D_{\mathrm{KL}}\left(L(\theta|\mathbf{x})||L(\theta+\Delta\theta|\mathbf{x})\right). \tag{216}
$$

11.4.2 Multidimensional Fisher Information

The ideas on Fisher information in §11.4.1 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

$$
\mathbb{I}_F\mathbf{X}_{\boldsymbol{\theta}} \equiv -\mathbb{E}_{\boldsymbol{\theta}}\nabla^2\ln f(\mathbf{X}|\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\begin{pmatrix} \frac{\partial^2}{\partial\theta_1^2}\ln f(\mathbf{X}|\boldsymbol{\theta}) & \cdots & \frac{\partial^2}{\partial\theta_1\partial\theta_m}\ln f(\mathbf{X}|\boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\theta_m\partial\theta_1}\ln f(\mathbf{X}|\boldsymbol{\theta}) & \cdots & \frac{\partial^2}{\partial\theta_m^2}\ln f(\mathbf{X}|\boldsymbol{\theta}) \end{pmatrix} \tag{217}
$$

for which the univariate second-order derivative in (212) 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,

$$
\ln L(\boldsymbol{\theta}+\Delta\boldsymbol{\theta}|\mathbf{x}) \approx \ln L(\boldsymbol{\theta}|\mathbf{x}) + \Delta\boldsymbol{\theta}^\top\ln L(\boldsymbol{\theta}|\mathbf{x}) + \frac{1}{2}\Delta\boldsymbol{\theta}^\top H\ln L(\boldsymbol{\theta}|\mathbf{x})\Delta\boldsymbol{\theta} \tag{218}
$$

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

$$\mathbb{H}\mathbf{X}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}} \approx \mathbb{H}\mathbf{X}_{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta}^{\top}\frac{1}{2}\mathbb{I}_F\mathbf{X}_{\theta}\Delta\boldsymbol{\theta}. \tag{219}$$

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

$$D_{\mathrm{KL}}\left(L(\boldsymbol{\theta}|\mathbf{x})||L(\boldsymbol{\theta}+\Delta\boldsymbol{\theta}|\mathbf{x})\right) = \Delta\boldsymbol{\theta}^{\top}\frac{1}{2}\mathbb{I}_F\mathbf{X}_{\boldsymbol{\theta}}\,\Delta\boldsymbol{\theta} \tag{220}$$

supplies the value of the second-order correction,

$$\mathbb{H}\mathbf{X}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}} \approx \mathbb{H}\mathbf{X}_{\boldsymbol{\theta}} + D_{\mathrm{KL}}\left(L(\boldsymbol{\theta}|\mathbf{x})||L(\boldsymbol{\theta}+\Delta\boldsymbol{\theta}|\mathbf{x})\right). \tag{221}$$

12 Bayesian Perspectives

12.1 Bayes' Theorem

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

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{B|A\}\mathbb{P}\{A\}}{\mathbb{P}\{B\}} \tag{222}$$

12.2 Parameter Refinement and Estimation

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

12.3 Conjugate Families

Beta prior, binomial likelihood (n draws):

$$\left.\begin{array}{ll} \text{beta prior:} & \mathrm{B}(\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ \text{binomial likelihood:} & \mathrm{Bin}(n,k) \propto \theta^{k}(1-\theta)^{n-k} \end{array}\right\} \Rightarrow \text{beta posterior: } \mathrm{B}(\alpha+k, \beta+n-k) \tag{223}$$

Gamma prior, Posson likelihood (n draws):

$$\left.\begin{array}{ll} \text{gamma prior:} & \Gamma(\alpha,\beta) \propto x^{\alpha-1}e^{-\beta x} \\ \text{Poisson likelihood:} & \mathrm{Poi}(x) \propto x^{k}e^{-nx} \end{array}\right\} \Rightarrow \text{gamma posterior: } \Gamma(\alpha+k, \beta+n) \tag{224}$$

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

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

$$
\Rightarrow \text{Gaussian posterior: } \mu \sim \mathrm{N}\left(\frac{\kappa_0\mu_0 + n\bar{\mathbf{x}}}{\kappa_0 + n}, \frac{\sigma^2}{\kappa_0 + n}\right) \tag{225}
$$

$$
\kappa_0 = \frac{\sigma}{\sigma_0} \tag{226}
$$

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

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

$$
\Rightarrow \text{inverted gamma posterior: } \sigma^2 \sim \mathrm{I}\,\Gamma\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \tag{227}
$$

12.4 Monte Carlo Methods