Expanding Man's Probability Review

Expanding Man

Abstract
A review of the fundamentals of probability theory.

Contents

1 Introduction

Probability, in addition to being fundamental to most areas of science, has immediate relevance to everyday life. Despite this, probability theory did not see significant development until the late renaissance, later even than differential calculus, which is perhaps a reflection of its reputation for being subtle and unintuitive.

In this review I will give a self-contained description of probability theory and some of its simplest applications.

2 Fundamentals

2.1 The Probability Function

Probability is defined over a set called the **sample set** (sometimes referred to as a **sample space**). The elements of this set can be interpreted as "events" or "possibilities". For example, the sample set of flipping a coin is $\{\text{heads}, \text{tails}\}$; for rolling a cubical die it's $\mathbb{Z}_6$; for `2d20` it's $\mathbb{Z}_{20} \times \mathbb{Z}_{20}$. For a sample set

- The elements should be *mutually exclusive* in some sense.
- *All possible outcomes* should be contained in the set.

We define the **probability function** over a sample set Ω

$$\Pr : 2^\Omega \to [0, 1] \tag{1}$$

2^Ω is a somewhat fanciful notation for the power set of Ω (i.e. $2^\Omega = \{A \mid A \subseteq \Omega\}$).[1] $[0, 1]$ is the interval on $\mathbb{R}$ from 0 to 1, inclusive of the endpoints.

To "**sample**" Ω means to select an element from it stochastically, such that the probability to select $\omega \in \Omega$ is $\Pr(\{\omega\})$. This often coincides with some intuitive notion of sampling, such as the result of some physical process, such as flipping a coin, though the general concept is more abstract and not contingent on such a process. A common way to describe the probability of a finite sample set is with the "urn" analogy. For example, $\Pr(\{\omega\})$ is the probability to draw a marble with label ω from an urn.

A function must satisfy reasonable constraints to be interpretable as a probability function. For $A, B \subseteq \Omega$

$$\begin{aligned} &\Pr(\emptyset) = 0 \\ &\Pr(\Omega) = 1 \\ A \cap B = \emptyset \implies &\Pr(A \cup B) = \Pr(A) + \Pr(B) \end{aligned} \tag{2}$$

The first of these can be interpreted as the statement that sampling Ω results in some $\omega \in \Omega$. The second condition states that we are certain to sample one of *any* elements of the sample set and serves as a normalization condition for Pr.

The last property in (2) states that we can add the probabilities of disjoint sets. This lets us derive some further properties. In the remainder of this section, let $A, B \subseteq \Omega$. From the axioms of set theory

$$A = (A \setminus B) \cup (A \cap B) \tag{3}$$

Combining this with the third line of (2) we find

$$\Pr(A) = \Pr(A \setminus B) + \Pr(A \cap B) \tag{4}$$

From this and the positive definiteness of Pr, it immediately follows that

$$\Pr(A) \geq \Pr(A \cap B) \tag{5}$$

Also, since $A \cup B = (A \setminus B) \cup B$ we have

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) \tag{6}$$

Note that this implies a "triangle inequality"

$$\Pr(A) + \Pr(B) \geq \Pr(A \cup B) \tag{7}$$

It is sometimes said that "the probability of A or B is $\Pr(A) + \Pr(B)$", here we have seen that this is only true of disjoint sets (in particular singleton sets of distinct elements, which is the statement this usually refers to).

Sometimes we will refer to the sample set as a **statistical ensemble**. This is a term common in physics. For example the $6N$-dimensional phase space of N particles in a box can also be thought of as a statistical ensemble,

[1] This notation comes from the fact that $\left|2^\Omega\right| = 2^{|\Omega|}$, that is 2^Ω has $2^{|\Omega|}$ elements.

where the probability is of each point in the phase space. We will make it more clear what this means in the case of a continuous set in the next section.

2.2 Probability Measure

In this note we adopt the so-called measure-theoretic definition of probability, albeit somewhat informally. The meaning of a probability function Pr perhaps seems intuitively obvious for discrete sets, but requires more exposition for cases in which Ω is continuous.

For this we must introduce the concept of a **measure**. For our purposes a measure is a function $\mu : 2^\Omega \to \mathbb{R}_\geq$ such that $\mu(\emptyset) = 0$. We define a **probability measure** P such that

$$\Pr(A) = \int_A \mathrm{d}P \tag{8}$$

At this point the definition constitutes not much more than a change in notation defining integrals with respect to some differential $\mathrm{d}P$. Eventually this will allow us to deploy the machinery of integral calculus to probabilities on continuous sets.

From the properties of Pr it then follows that $\int_\Omega \mathrm{d}P = \Pr(\Omega) = 1$, while $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ for $A \cap B = \emptyset$ follows from the usual axioms of integration, which we will not detail here. (8) is more useful if we wish to compute probabilities given some coordinate system $x : \Omega \to \mathbb{R}^n$ for which we can define

$$\mathrm{d}P(x) = p(x)\,\mathrm{d}x \tag{9}$$

where $p(x)$ is known as a **probability density**. It follows trivially from what we have already discussed that $p(x) \geq 0$, $\forall x$ and $\int \mathrm{d}x\, p(x) = 1$.

The density $p(x)$ is often referred to as a probability density *function* but we should emphasize that it need not be a function in the formal sense since we only require that integrals of it over finite subsets of Ω be defined, and not necessarily that $p(x)$ is finite or well-defined for all x. An important example of a density which is manifestly not a function is the ironically named **Dirac δ-function**, which we can take as being defined by the condition

$$\int_A \mathrm{d}x\, f(x)\,\delta(x - x_0) = \begin{cases} f(x_0) & \Leftarrow x_0 \in A \\ 0 & \text{else} \end{cases} \tag{10}$$

Clearly $\delta(x - x_0)$ must be appropriately normalized $\int \mathrm{d}x \delta(x) = 1$ to serve as a valid probability density. As a probability density, the Dirac δ means, roughly "certain to be at the point x_0".

2.3 Cumulative Probability

An important special case is when the coordinate is one dimensional $x : \Omega \to \mathbb{R}$, in which we refer to the probability as **univariate**. In such cases we define the **cumulative distribution function** which is the anti-derivative of the probability density

$$F(x) \coloneqq \Pr(\xi \leq x) = \int_{-\infty}^{x} \mathrm{d}\xi\, p(\xi) \tag{11}$$

[2] From this it follows that

$$\Pr(a \leq x \leq b) = F(b) - F(a) \tag{12}$$

There's not much to say about the cumulative probability in and of itself, but it is often a useful notational tool.

2.4 Multivariate Probability and Marginal Density

The general case where the sample set admits coordinates $x : \Omega \to \mathbb{R}^n$ with $n > 1$ is called **multivariate** probability. The definition of the probability density in this case is a straightforward extrapolation of (9). For example, in $\mathbb{R}^2$

$$\mathrm{d}P(x, y) = p(x, y)\,\mathrm{d}x\,\mathrm{d}y \tag{13}$$

In many applications we will be interested in integrating out some of the coordinates

$$\mathrm{d}P_X(x) \coloneqq \int \mathrm{d}y\,\mathrm{d}P(x, y) \tag{14}$$

where the integral is over the domain of y. $\mathrm{d}P_X$ is known as a **marginal probability measure**. Of course, we can define the density by $\mathrm{d}P_X(x) = p_X(x)\,\mathrm{d}x$ where $p_X(x)$ is known as a **marginal density**.

[2] The use of the symbol F here may seem capricious, but it is commonly used. Often a probability density p is written f.

It should be obvious that $\mathrm{d}P_X$ is itself a valid probability measure. Intuitively, we interpret it as "the probability of x regardless of y". Because of this, in many contexts the distinction between probability and marginal probability is implied only weakly.

2.5 Conditional Probability

What is the probability that $x \in A$ on the condition that $x \in B$? The preceding discussion does not make it entirely clear what this means. To address this, we must define **conditional probability** $\Pr(A|B)$ (read "the probability of A such that B). Let's impose some requirements on this definition. First, we'd like

$$\Pr(A|B) \propto \Pr(A \cap B) \qquad \forall A, B \subseteq \Omega \tag{15}$$

i.e. that the conditional probability is always proportional to the probabilty over Ω. Next, we require

$$\Pr(B|B) = 1 \tag{16}$$

to coincide with our intuitive notion of conditioning (i.e. we must always have $x \in B$ if we impose $x \in B$ as a constraint). We therefore define

$$\Pr(A|B) \coloneqq \frac{\Pr(A \cap B)}{\Pr(B)} \tag{17}$$

It is not in general possible to define a fully self-consistent conditional probability measure. The root issue is that, generically, the denominator of (17) for infinitessimal events vanishes, causing the would-be measure to diverge. The best we can do is to define, for example

$$\mathrm{d}P(x \,|\, y \in B) = \frac{\mathrm{d}x}{\Pr(B)} \int_B \mathrm{d}y \, p(x, y) \tag{18}$$

which is only valid when $\Pr(B) = \int \mathrm{d}x \int_B \mathrm{d}y \, p(x, y) > 0$. We should caution that this definition does not seem to be universally recognized.

The reader should be aware that a great deal of literature, particularly for machine learning, will casually introduce conditional densities written $p(x \,|\, y)$. Often enough careful thought will reveal problems with these. In practice, the most important property of such objects is that they can be used to consistently define marginal densities. For example

$$p(x) = \int \mathrm{d}y \, p(x \,|\, y) p(y) \tag{19}$$

which requires that

$$p(x \,|\, y) = \frac{p(x, y)}{p(y)} \tag{20}$$

at least in regions where $p(y) > 0$. On the other hand, these objects may not have any obvious relationship to the true conditional probabilities, so caution is advised.

2.6 Random Variables

When dealing with multivariate probability densities, it is often convenient to introduce the concept of **random variables**. A random variable is a map $X : \Omega \to \mathcal{M}$ where $\mathcal{M}$ is an n-dimensional differentiable manifold (which we will usually take to be $\mathbb{R}^n$). Random variables are mostly a notational convenience that make it easier to talk about more complex relationships between elements of a sample set, and they aren't really any different from the coordinates on Ω we've already been discussing. For example, let's define $f : \mathbb{R}^2 \to \mathbb{R}$ and write

$$Z = f(X, Y) \tag{21}$$

where X, Y, Z are random variables. By definition, the probability measure for Z is

$$\mathrm{d}P_Z(z) = \mathrm{d}z \int \mathrm{d}P_{XY}(x, y) \, \delta(z - f(x, y)) \tag{22}$$

We will sometimes use the notation $X \sim p$ to mean that p is the probability density for the random variable x, read "X is distributed according to p".

We say that two random variables are **independent** if their measures can be factorized

$$\mathrm{d}P_{XY}(x, y) = \mathrm{d}P_X(x) \, \mathrm{d}P_Y(y) \tag{23}$$

The normalization conditions ensure that (23) is the unique factorization of $\mathrm{d}P_{XY}$ (i.e. the factors are always the marginal measures) when X and Y are independent.

Two random variables are **independent and identically distributed** (abbreviated **i.i.d.**) if in addition to (23) $p_X(\xi) = p_Y(\xi)$.

If we write an expression that relates multiple random variables, it must be possible to define them over the same sample set so that they share an overall distribution called the **joint distribution**. For example, if we define two random variables X, Y, for any function $f(X, Y)$ to make sense, it must be possible to define $\mathrm{d}P_{XY}(x, y)$. Accordingly, X and Y are independent iff $p_{XY}(x, y) = p_X(x)\, p_Y(y)$.

2.6.1 Example: Sum of Two Independent Random Variables

Consider $Z = X + Y$ where X and Y are independent. According to our definition

$$p_Z(z) = \int \mathrm{d}x\, \mathrm{d}y\, \delta(z - x - y)\, p_X(x)\, p_Y(y) \tag{24}$$

Integrating over one of the variables we have

$$p_Z(z) = \int \mathrm{d}x\, p_Y(z - x)\, p_X(x) \tag{25}$$

Note that we are treating it as implicit that $p_Z(z) = 0$ for any z which is not expressible as $z = x + y$, which is important if, for example, x or y have a finite range.

(25) states that the density of the sum of two random variables is the convolution of their densities. We will repeatedly find this important when discussing prominent examples of probability densities.

2.7 Change of Coordinates

One of the useful consequences of defining probability as a measure is that it implicitly comes with the machinery of differential geometry. It is true, for our purposes, by definition, that under a change in coordinates $x \to y$

$$\begin{aligned} \mathrm{d}P(y) &= \mathrm{d}P(x) \\ p'(y)\, \mathrm{d}y &= p(x)\, \mathrm{d}x \end{aligned} \tag{26}$$

Therefore

$$p'(y) = p(x) \left| \frac{\partial y}{\partial x} \right|^{-1} \tag{27}$$

where $\left|\frac{\partial y}{\partial x}\right|$ is the Jacobian determinant of y with respect to x. We will not attempt to rigorously justify (27), but it is derivable with the tools of differential geometry and measure theory. Fortunately it is also intuitive based on remedial calculus alone.

2.7.1 Probability as a Differential Form

This section requires some knowledge of the exterior calculus and the language of differential forms.

If the sample space is an n-dimensional manifold, $\mathrm{d}P$ must be an n-form, which is unique up to an overall factor. Therefore

$$\mathrm{d}P = p(x_1, x_2, ..., x_n)\, (\mathrm{d}x_1 \wedge \mathrm{d}x_2 \wedge \cdots \wedge \mathrm{d}x_n) \tag{28}$$

Clearly (27) now follows from the standard transformation properties of the basis n-forms. On manifolds equipped with a metric, we must include a geometric factor in explicit coordinate representations of the integration over $p(x)$

$$\Pr(A) = \int_A \mathrm{d}^n x\, \sqrt{\pm g}\, p(x) \tag{29}$$

At our level of rigor, differential forms are essentially measures[3], so this is consistent with our discussion in the last section.

2.7.2 Example: Rotations

Consider the transformation

$$x \mapsto R(\varphi)\, x \tag{30}$$

[3]Technically there is an important distinction involving orientation which is well beyond the scope of this review.

where x is a vector and $R(\varphi)$ is a rotation matrix. Such matrices form a group, specifically $\mathrm{SO}(N)$. One of the properties of a group is that every element must have an inverse, in our caes $R^{-1}(\varphi)R(\varphi) = 1$. For $x' = R(\varphi)\,x$ we therefore have

$$p'(x') = p(x)\left|\frac{\partial x}{\partial x'}\right| = p\big(R^{-1}(\varphi)x'\big) \tag{31}$$

since $\left|R^{-1}\right| = 1$ is a property of $\mathrm{SO}(N)$.

2.8 Bayes' Lemma

We now come to a very famous but trivial result. By combining (17) for $\Pr(A|B)$ and $\Pr(B|A)$, we find

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \tag{32}$$

One reason that this result is so famous is that it is very common for the probabilities of some set not to be directly observable. That is, you may want to know $\Pr(A|B)$, but can only observe $\Pr(B|A)$. (32) relates these.

The factor $\Pr(A)$ in the numerator of the right hand side of (32) is commonly referred to as a **prior**. As we will discuss, in many applications it is not directly observable and must be postulated. The denominator $\Pr(B)$ can be thought of as a normalization factor, and is often determined by requiring $\Pr(\Omega) = 1$.

2.9 Epistemology

Most of the above discussion has been rather abstract, but to apply probability to the real world, we must resort to some discussion of interpretation. We caution the reader that interpretation of probability theory is an enormous subject in and of itself, in both the realms of pure mathematics and philosophy, but here our goal is to avoid the proverbial rabbit hole and provide the bare minimal context required for a useful working knowledge of probability.

There are primarily two equivalent interpretations of probability theory. As a common premise, imagine that we have prepared some number n of *identical* experiments. The word identical is doing a lot of work here, but for now we decline a more detailed discussion of what it means. For our purposes it will suffice to assume we have formulated some reasonable definition. The ith experiment concludes with an observation $\omega_i \in \Omega$, where Ω is the sample space and the set of all possible outcomes of the experiment.

The **frequentist** interpretation asserts

$$\Pr(A) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I(\omega_i \in A) \qquad\qquad \forall A \subseteq \Omega \tag{33}$$

where I is an indicator that gives 1 if its argument is true and 0 otherwise. Thus, in this interpretation, the probability of A is simply the ratio of events in A to the total number of events as that number approaches infinity. Of course, the statement $\Pr(A) = a_0$ does not necessarily mean that it has been observed infinitely many times, but that we could reject $\Pr(A) = a_0$ as a plausible statement if repeated observations fail to converge to this value.

The other prominent interpretation of probability theory is called the **Bayesian** interpretation. It is easier to describe it by imagining that we have some theory the parameters of which we will collectively refer to as θ. Bayesian probability treats θ as a collection of random variables. Then, by definition

$$P(\theta|(\omega_1, \omega_2, ..., \omega_n)) = \frac{P((\omega_1, \omega_2, ..., \omega_n)|\theta)P(\theta)}{P((\omega_1, \omega_2, ..., \omega_n))} \tag{34}$$

where $(\omega_1, \omega_2, ..., \omega_n)$ is a sequence of observations. Unobservability of the denominator seems like a problem, but as we have discussed it can be thought of as a normalization factor. $P((\omega_1, ..., \omega_n)|\theta)$ should be possible to compute from the theory and can be tought of as a requisite for a valid theory. The other factor in the numerator, the prior $P(\theta)$, poses a more serious and fundamental challenge. One way to proceed would be to choose it as a postulate, but in so doing we risk choosing a "pathological" prior, for example a delta function. Such pathological priors could prevent $P(\theta|(\omega_1, ..., \omega_n))$ from converging to an empirically plausible result.

The relative merits of each of these two interpretations are still the subject of much debate. In my opinion, the frequentist interpretation is simpler and comes with significantly less "conceptual baggage", but these interpretations are nevertheless equivalent. Furthermore, the procedure implied by the Bayesian interpretation, which that interpretation treats as fundamental (i.e. using the axioms of probability theory to update assumptions

with new evidence) are of great practical importance even if one takes the frequentist view entirely literally. It is useful to maintain a pluralistic mindset in regard to this subject, as it is for many others.

3 Statistics

A statistic is a functional of a probability density. In a sense, they describe the infinite dimensional space of all possible probability densities in finite dimensional terms.

3.1 Expectation Value

Arguably the most important, and certainly the most common type of statistic is known as an **expectation value**. For a random variable X it is defined by

$$\langle f(X) \rangle = \int_\Omega \mathrm{d}P(x)\, f(x) \tag{35}$$

On a discrete set we can write

$$\langle f \rangle = \sum_{A_j \in \Omega} f(A_j) \Pr(A_j) \tag{36}$$

where the A_j are elements of a partition of Ω, that is $\bigcup_j A_j = \Omega$, $A_i \cap A_j = \emptyset, \forall i \neq j$.

The variety of notations available for expectation values is bewildering. Here we will use $\langle \cdot \rangle$, which is most common among physicists. Mathemeticians will often use $E(\cdot)$, $E[\cdot]$, or $\mathbb{E}[\cdot]$. We will adopt the convention of writing expressions inside the $\langle \cdot \rangle$ as functions of random variables, such that the expectation is an integral over densities of those variables. In contexts where it may be confusing what we are integrating over, we will write the name of the random variable being integrated over in the subscript, e.g. $\langle \cdot \rangle_X$.

It should be obvious that for any constant c, $\langle c \rangle = c$. The simplest non-trivial expectation value is then $\langle X \rangle$, which is called the **mean**. That is $\langle X \rangle$ is the mean of the probability density $p_X(x)$.

Some simple examples are in order. For $X \sim \delta(x - x_0)$, we have $\langle X \rangle = x_0$, since no other value is supported by the density. For a "flat" distribution

$$p(x) = \begin{cases} 1 & \Leftarrow x \in [0,1] \\ 0 & \Leftarrow x \notin [0,1] \end{cases} \tag{37}$$

we compute

$$\langle X \rangle = \int_0^1 \mathrm{d}x\, x = \left. \frac{x^2}{2} \right|_{x=1} = \frac{1}{2} \tag{38}$$

which conforms with the intuitive notion of the value "in the middle" of p_X. It should be obvious, but we nonetheless emphasize, that $\langle X \rangle$ *need not* be a "likely" value of X in any sense. For example, for $X \sim \frac{\delta(x) + \delta(x-1)}{2}$ we have $\langle X \rangle = \frac{1}{2}$, but we cannot sample any values in a neighborhood of $\frac{1}{2}$. $\langle X \rangle$ is often interpreted as providing a "typical value" of X, though, as we have seen, this interpretation has limitations depending on the distribution. If the distribution has a very long tail, $\langle X \rangle$ can be quite an *atypical* value since the distribution converges only very slowly for large x.

We can sompute $\langle f(X) \rangle$ for any function f for which the integral converges, but some of these statistics are so commonly used that they have special names. The **variance** is

$$\mathrm{var}[X] \coloneqq \left\langle (X - \langle X \rangle)^2 \right\rangle \tag{39}$$

which we can simplify to

$$\mathrm{var}[X] = \langle X^2 \rangle - 2\langle X \langle X \rangle \rangle + \langle X \rangle^2 = \langle X^2 \rangle - \langle X \rangle^2 \tag{40}$$

Sometimes the symbol σ^2 is associated with the variance, for example we may write $\sigma_X^2 = \mathrm{var}[X]$. $\sigma_X = \sqrt{\mathrm{var}[X]}$ is called the **standard deviation**. For Gaussian-like distributions (which we will introduce in detail in Section 4.1), σ can be thought of as describing the "width" of the distribution.

Similarly

$$m_n = \frac{\left\langle (X - \langle X \rangle)^n \right\rangle}{\sigma_X^n} \tag{41}$$

are known as **moments** of the distribution of X. m_3 is also called the **skewness** and m_4 is also called the **kurtosis**.

3.2 Mode

A **mode** is a point at which a distribution reaches its supremum

$$\arg\sup_x p(x) \tag{42}$$

Distributions with a unique mode are called **unimodal**, those with exactly two distinct modes are **bimodal**. Of course the modes can be computed by solving $\partial_x p(x) = 0$ and checking which results are maxima.

3.3 Quantiles

Quantiles, along with the mode, are arguably the only commonly used statistics which are not expressible as expectation values. They are only relevant in cases where Ω is isomorphic with $\mathbb{R}$, in which sense they are univariate statistics only. Consider

$$\frac{1}{q} = \int_\xi^\infty \mathrm{d}x\, p_X(x) \tag{43}$$

The quantile in this expression is ξ, which is the value such that "the rest of the distribution after ξ" makes up $1/q$ of the total probability. More generally, we can define a set $\{\xi_0, \xi_1 ..., \xi_q\}$ such that

$$\frac{1}{q} = \int_{\xi_k}^{\xi_{k+1}} \mathrm{d}x\, p_X(x) \tag{44}$$

That is, we can partition the interval on $\mathbb{R}$ in which p_X has support into q parts, each with an equal total probability of $1/q$. As an example, a 4-quantile is also called a **quartile**, using (37) restricted to $[0, 1]$ as an example we have

$$\xi_0 = 0 \qquad \xi_1 = \frac{1}{4} \qquad \xi_2 = \frac{1}{2} \qquad \xi_3 = \frac{3}{4} \qquad \xi_4 = 1 \tag{45}$$

Alternatively, if we continue $p_X(x)$ to all of $\mathbb{R}$, keeping its value at 0 outside of $[0, 1]$, we'd have $\xi_0 = -\infty$ and $\xi_4 = \infty$. The 0th and qth q-quantiles are always the minimum or maximum values on which the distribution has support, and can always be taken to be $\pm\infty$, for which reason they are usually not referred to. This means that there are always $q - 1$ non-trivial q-quantiles. In terms of the cumulative probability F_X we have

$$\frac{1}{q} = F_X(\xi_{k+1}) - F_X(\xi_k) \tag{46}$$

There is only a single finite 2-quantile and this is referred to as the **median**. It's worth noting that the median is often used along with or in place of the mean as indicating a "typical value", but it is less sensitive to outliers in the sense that a distribution which falls off only slowly toward $\pm\infty$ will tend to have a much larger (as in much more positive or much more negative) mean than median.

If the sample set Ω has some concept of ordering we can define quantiles also for discrete sets, but we cannot in general guarantee that all of the quantiles are defined. For example, consider the set $\{0, 1, 2\}$, with $\Pr(0) = \frac{1}{2}$, $\Pr(1) = \Pr(2) = \frac{1}{4}$. Here 1 is the median since $\Pr(\{x \mid x \geq 1\}) = \Pr(1) + \Pr(2) = \frac{1}{2}$. 2 is the largest quartile since $\Pr(\{x \mid x \geq 2) = \Pr(2) = \frac{1}{4}$. There are however, no 3-quantiles, since there is no $\xi \in \{0, 1, 2\}$ such that $\Pr(\{x | x \geq \xi\}) = \frac{1}{3}$. This can only occur for discrete sets or discontinuous probability densities. All q-quantiles for all $q \geq 2$ exist for continuous densities.

Some fields of study make frequent use of 100-quartiles which are also known as **percentiles**.

As we have previously mentioned, the generalization of quantiles to higher dimensions is not straightforward. A quantile in n dimensions is an $n - 1$ dimensional hypersurface, but it is also not in general unique. We can, however, talk about quantiles of 1-dimensional marginal probability distributions. For example, if we have some density $p_{XY}(x, y)$, we can discuss quantiles of p_X and p_Y, that is, we let one of the coordinates define the hypersurfaces which we take as the quantiles.

3.4 Multivariate Statistics

In most of the preceding discussion we have avoided explicitly specifying whether Ω has one or more dimensions. Indeed, the generalization of (35) to multiple dimensions is trivial. Nevertheless, there are some common notations, conventions and terminology specific to multivariate statistics, so we will review them here. We will

denote multidimensional random variables with an index, e.g. X^i, which unless otherwise specified we will take to be vectors on $\mathbb{R}^n$.

The obvious generalization of the mean is

$$\langle X^i \rangle = \int \mathrm{d}^n x \, x^i \, p_X(x) \tag{47}$$

The probability density p_X here is no different than more explicitly multivariate densities we have already seen such as $p_{XY}(x, y)$ except with a more compact notation. We will often write $\mu_X^i \coloneqq \langle X^i \rangle$, or simply μ^i for convenience, when there is no risk of confusion.

While the generalization of the mean to higher dimensions is a vector, the generalization of the variance is a symmetric matrix

$$\Sigma_X^{ij} \coloneqq \langle (X^i - \langle X^i \rangle)(X^j - \langle X^j \rangle) \rangle \tag{48}$$

This is referred to as the **covariance** matrix. When written in terms of two scalar random variables X and Y, this is often written

$$\mathrm{cov}[X, Y] \coloneqq \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle \tag{49}$$

which is of course a scalar. We can extend to higher dimensional X and Y

$$\mathrm{cov}[X^i, Y^\alpha] = \langle (X^i - \langle X^i \rangle)(Y^\alpha - \langle Y^\alpha \rangle) \rangle \tag{50}$$

where we have written X and Y with different types of indices to emphasize that they needn't be of the same number of dimensions to define the covariance. In such a case the covariance matrix is non-square. $\Sigma_X^{ij} = \mathrm{cov}[X^i, X^j]$ is sometimes called an **auto-covariance** matrix.

The **correlation** coefficient, is simply the covariance scaled by the standard deviation of each variable, for example

$$\mathrm{corr}[X, Y] \coloneqq \frac{\mathrm{cov}[X, Y]}{\sigma_X \sigma_Y} \tag{51}$$

which is only worth mentioning because it is frequently referred to.

Let's take a moment to examine how the elements of Σ^{ij} should be interpreted. Clearly the diagonal elements are merely the variances of each component. The off-diagonal terms can be written

$$\mathrm{cov}[X, Y] = \langle XY \rangle - \mu_X \langle Y \rangle - \mu_Y \langle X \rangle + \mu_X \mu_Y = \langle XY \rangle - \mu_X \mu_Y \tag{52}$$

We know that

$$\langle XY \rangle = \int \mathrm{d}P_{XY}(x, y) \, xy \tag{53}$$

When X and Y are independent we can factorize this into

$$\left(\int \mathrm{d}P_X(x) \, x \right) \left(\int \mathrm{d}P_Y(y) \, y \right) = \mu_X \mu_Y \tag{54}$$

in which case $\mathrm{cov}[X, Y] = 0$. In the other limiting case, where $X = Y$, we have $\mathrm{cov}[X, Y] = \langle X^2 \rangle - \mu_X^2 = \mathrm{var}[X] = \mathrm{var}[Y]$. Therefore, random variables with independent components have diagonal covariance matrices, and as we increase their correlation we approach a uniform matrix with all elements equal to $\mathrm{var}[X]$.

3.5 Characteristic Functions

The **characteristic function** of a random variable X is the Fourier transform of its probability density

$$\phi_X(k) = \int \mathrm{d}P_X(x) \, e^{ikx} = \langle e^{ikX} \rangle \tag{55}$$

Basic facts about Fourier transforms tell us that this can be inverted

$$p_X(x) = \int \frac{\mathrm{d}k}{2\pi} e^{-ikx} \phi_X(k) \tag{56}$$

Characteristic functions are useful mainly as an analytical tool the same way that Fourier transforms are useful more generally. As an important example, consider the sum of two random variables $Z = X + Y$. As we have seen in (25), the probability density of Z is the convolution of those of Y and X. We can then write

$$\begin{aligned} p_Z(z) &= \int \mathrm{d}x \int \frac{\mathrm{d}k}{2\pi} \int \frac{\mathrm{d}l}{2\pi} e^{-ilz} e^{ilx} e^{-ikx} \phi_X(k) \phi_Y(l) \\ &= \int \frac{\mathrm{d}l}{2\pi} e^{-ilz} \left(\int \mathrm{d}x \int \frac{\mathrm{d}k}{2\pi} e^{i(l-k)x} \phi_X(k) \phi_Y(l) \right) \end{aligned} \tag{57}$$

From this and the definition (55), we can simply read off the characteristic function for Z

$$\phi_Z(l) = \int \mathrm{d}x \int \frac{\mathrm{d}k}{2\pi} e^{i(l-k)x} \phi_X(k) \phi_Y(l) \tag{58}$$

Recognizing that the Fourier transform of a plane wave is the δ function

$$\delta(x) = \int \frac{\mathrm{d}k}{2\pi} e^{ikx} \tag{59}$$

we find

$$\phi_Z(l) = \int \mathrm{d}l \, \delta(k-l) \phi_X(k) \phi_Y(l) = \phi_X(l) \phi_Y(l) \tag{60}$$

More succinctly, for posterity

$$\phi_{X+Y}(k) = \phi_X(k) \phi_Y(k) \tag{61}$$

That is, the characteristic of the sum of random variables is the product of their characteristic functions. This should be familiar to anyone used to Fourier transforms as the convolution theorem, which says essentially the same thing in somewhat different language.

The above can be trivially generalized to arbitrarily many variables, and to a general linear combination of those variables by re-scaling

$$\phi_{a_1 X_1 + \cdots + a_n X_n} = \phi_{X_1}(a_1 k) \cdots \phi_{X_n}(a_n k) \tag{62}$$

Note also that from $\varphi_X(k) = \langle e^{ikX} \rangle$, we can expand the exponential. We can always shift a random variable to have zero mean by translation $X \to X - \mu$ so that

$$\varphi_X(k) = \langle e^{ikX} \rangle = e^{ik\mu} \left\langle \sum_{n=0}^{\infty} \frac{i^n k^n (X-\mu)^n}{n!} \right\rangle = e^{ik\mu} \sum_{n=0}^{\infty} \frac{i^n k^n \sigma^n m_n}{n!} \tag{63}$$

where $m_n = \langle (X-\mu)^n \rangle / \sigma^n$ is the nth moment. Note that the $n = 1$ term always vanishes, by construction. We could have redefined $X \to X' = \frac{X-\mu}{\sigma}$ to eliminate both the phase $e^{ik\mu}$ and the σ, which is sometimes convenient. We will see in our discussion of some specific distributions that this expansion is sometimes useful.

4 Distributions

Here we discuss the properties of certain specific classes of probability density function. As we will see, one of these in particular, the Gaussian or normal distribution, acts as a kind of limiting case of all other distributions, in a sense that we will see below. Many of the other most important examples of specific classes of distributions are formed by combining Gaussians in some way.

4.1 Gaussian Distribution

Name	Parameters	Support	Mean	Variance
$\mathcal{N}(\mu, \sigma^2)$	$\mu \in \mathbb{R}, \sigma > 0$	$\mathbb{R}$	μ	σ^2

The **Gaussian** or **normal** distribution is uniquely important in a sense we will describe below. The univariate Gaussian distribution has the form

$$X \sim \mathcal{N}(\mu, \sigma) \implies \qquad p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{64}$$

where $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$ are parameters. Their names are not coincidental, since

$$\begin{aligned}\langle X\rangle &= \mu \\ \mathrm{var}[X] &= \sigma^2\end{aligned} \tag{65}$$

The factor $\left(\sigma\sqrt{2\pi}\right)^{-1}$ should be thought of as simply a normalization factor for ensuring $\int \mathrm{d}x\, p_X(x) = 1$. Note that the Gaussian is symmetric about $x = \mu$. As a consequence, its skewness, and all of its odd-numbered moments vanish.

Computing the cumulative probability of a Gaussian is non-trivial. Due to the importance of the distribution, a transcendental function called the **error function** is defined

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{d}\xi\, e^{-\xi^2} \tag{66}$$

This is simply an integral over half the Gaussian after a change of variables. This function is the subject of a great deal of study in its own right, and its properties are extensively documented. The cumulative probability is therefore

$$F_x(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \tag{67}$$

Arguably the property of this distribution that makes it so simple is that the characteristic function of a Gaussian is also a Gaussian up to a normalization factor

$$\phi(k) = e^{ik\mu} e^{-\frac{1}{2}\sigma^2 k^2} \tag{68}$$

From this and (61) we can conclude that for two random variables $X, Y \sim \mathcal{N}(\mu, \sigma)$

$$Z = X + Y \implies \qquad Z \sim \mathcal{N}\left(\mu, \sqrt{2}\sigma\right) \tag{69}$$

This is trivially generalized to the case of arbitrarily many independent i.i.d. random variables. This is an exact special case of the more general central limit theorem which we discuss below.

4.1.1 Central Limit Theorem

The **Central Limit Theorem** (CLT) is arguably the single most important result in statistics, and as we will briefly discuss below one of the things that makes empirical science possible.

Let $\{X_1, X_2, ..., X_n\}$ be a set of i.i.d. random variables. A natural question is what is the distribution of their sum, that is, the distribution of

$$\overline{X}_n \coloneqq \frac{X_1 + X_2 + \cdots + X_n}{n} \tag{70}$$

We should expect that the distribution of $\overline{X}_n$ depends on the distribution of each of the X_j. Naively, we might also expect that this is true even in the limit $n \to \infty$. Remarkably, as we will show, this is not the case, instead the distribution of $\overline{X}_n$ is *always* a Gaussian, regardless of the distribution of each of the contituent variables.

To see how this occurs, we will make use of characteristic functions. Recalling the formula for the characteristic function of a linear combination of random variables (62), we write

$$\phi_{\overline{X}_n}(k) = \sqrt{n} \prod_{j=1}^{n} \phi_{X_j}\left(n^{-\frac{1}{2}}k\right) = \sqrt{n}\varphi_X^n\left(n^{-\frac{1}{2}}k\right) \tag{71}$$

where ϕ_X is the common characteristic function of all the X_j's. We have inserted the factor of $\sqrt{n}$ arbitrarily using (62) for reasons that will soon become apparently. Without loss of generality, we can choose each $\langle X_j\rangle = 0$ and $\mathrm{var}[X] = 1$ (since this is just a translation and re-scaling, it should be trivial to transform them back at the end of any calculation). Therefore

$$\phi_{\overline{X}_n}(k) = \sqrt{n}\left(1 - \frac{k^2}{2n} + \mathcal{O}\left(n^{-\frac{3}{2}}k^3\right)\right)^n = \sqrt{n}\left(1 - \frac{k^2}{2n}\right)^n + \mathcal{O}\left(n^{-\frac{3}{2}}k^3\right) \tag{72}$$

Using $e^x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n$, we find

$$\phi_{\overline{X}_n} = \sqrt{n}\, e^{-\frac{k^2}{2}} + \mathcal{O}\left(n^{-\frac{3}{2}}k^3\right) \tag{73}$$

As we have seen, the characteristic function of a Gaussian is itself a Gaussian, which is what we have here up to corrections of order $n^{-\frac{3}{2}}k^3$. So we have shown that

$$\lim_{n\to\infty} \frac{\sqrt{n}}{\sigma}\big(\overline{X}_n - \mu\big) \sim \mathcal{N}(0,1) \tag{74}$$

or, equivalently, after shifting by μ and scaling by $\frac{\sigma}{\sqrt{n}}$

$$\lim_{n\to\infty} \overline{X}_n \sim \mathcal{N}\big(\mu, \sigma/\sqrt{n}\big) \tag{75}$$

That is, the arithmetic mean of i.i.d. random variables, each with mean μ and σ^2 approaches a Gaussian with mean μ and variance σ^2/n. Crucially, this does *not* depend on the distribution of each variable in the sum.

Unsurprisingly, the mean of $\overline{X}_n$ is simply μ, the mean of each contributing variable. The variance of $\overline{X}_n$ however is *less* than the variance of each X_j. It's worth emphasizing that this is true of the *mean* $\overline{X}_n$, but *not* the sum $X_1 + \cdots + X_n$. An consequence of this with important practical implications is that, given any random process which we can sample freely, we can always construct a statistic with arbitrarily small variance. While $\overline{X}_n$ differs from a Gaussian by the fourier transform of the $\mathcal{O}\big(n^{-\frac{3}{2}}k^3\big)$ terms in the general case, in the case where the population distribution is also Gaussian, these correction terms vanish, and the distribution of the mean or sum is *always* a Gaussian with variance exactly $n\sigma^2$ for all values of n (and cosequently the variance of the mean is always σ^2/n).

The CLT plays a special role in science in that it allows us to obtain results that do not depend on detailed knowledge of probability distributions, in particular the uncertainty distributions of measurements. It also provides with a universal algorithm for reducing certain types of uncertainty: take more measurements. Since the variance of the mean is smaller than the variance of the population distribution, our confidence in the mena can be higher than our confidence in any individual measurement. This provides us with rigorous justification for the idea that taking more measurements reduces statistical uncertainties.

4.1.2 Multivariate Gaussian Distribution

Name	Parameters	Support	Mean	Variance
$\mathcal{N}_n(\mu, \Sigma)$	$\mu \in \mathbb{R}^n, \Sigma \in \mathbb{R}^n \times \mathbb{R}^n$	$\mathbb{R}^n$	μ	Σ

Generalizing $\mathcal{N}(\mu, \sigma^2)$ to higher dimensions is fairly trivial, but it is such an important case that we will take a moment to review it. We denote a set of random variables $X : \Omega \to \mathbb{R}^n$ that are distributed according to a multivariate Gaussian distribution as

$$X \sim \mathcal{N}_n(\mu, \Sigma) \tag{76}$$

where $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^n \times \mathbb{R}^n$ is the covariance matrix. This is precisely the covariance matrix we have already introduced $\Sigma^{ij} = \mathrm{cov}[X^i, X^j]$. It is left as an exercise for the reader to show that $X \sim \mathcal{N}_n(\mu, \Sigma) \Rightarrow \Sigma^{ij} = \mathrm{cov}[X^i, X^j]$. The probability density function can be obtained via normalization and is given by

$$\mathcal{N}_n(\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\!\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right) \tag{77}$$

Another useful fact about multivariate Gaussian variables is that the marginal distribution of the ith component variable is itself a Gaussian with variance Σ^{ii}. More generally, the marginal distribution for any subset of variables is simply the multivariate Gaussian with the variables in the complement of the subset deleted. For example, given variables $(X^1, X^2, X^3) \sim \mathcal{N}(\mu, \Sigma)$, the marginal distribution for (X^1, X^2) is simply $\mathcal{N}\big((\mu^1, \mu^2), \Sigma_{(12)}\big)$ where $\Sigma_{(12)}$ is Σ with the third row and column deleted. Showing this is tedious but can be done directly by integration.

Conditional distributions of the component variables are also Gaussian, but less straightforward. Suppose we partition $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ where $X_1 : \Omega \to \mathbb{R}^q$ and $X_2 : \Omega \to \mathbb{R}^{n-q}$. Accordingly we write

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad\qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{78}$$

Then $p_{X_1 \mid X_2}$ is a multivariate Gaussian with

$$\begin{aligned} \mu' &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \\ \Sigma' &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned} \tag{79}$$

As Gaussian distributions are so common in nature, thanks in no small part to CLT, the multivariate Gaussian distribution has many important applications, some of which we will explore later.

4.2 Γ Distribution

The class of 1-dimensional distributions which maximize entropy for a fixed $\langle X \rangle$ is called the Γ distribution. We will leave the formal introduction of entropy to sections on information theory, and focus on the χ^2 distribution, an important special case of Γ distribution.

4.2.1 χ^2 Distribution

Name	Parameters	Support	Mean	Variance
$\chi^2(k)$	$k \in \mathbb{Z}_>$	$\mathbb{R}_\geq$	k	$2k$

The χ^2 statistic is defined as

$$Q = \sum_{j=1}^{k} \frac{\left(X_j - \mu_j\right)^2}{\sigma_j^2} \tag{80}$$

where X_j are i.i.d. random variables, $\mu_j = \langle X_j \rangle$ and $\sigma_j = \mathrm{var}[X_j]$. In the special case where the $X_j \sim \mathcal{N}(\mu_j, \sigma_j)$, $Q \sim \chi^2(k)$ where $\chi^2(k)$ is known as the **χ^2 distribution** with k degrees of freedom. It is given by

$$p_Q(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma(k/2)} \qquad \forall x \geq 0 \tag{81}$$

where $\Gamma(x)$ is the Γ function ($\Gamma(n) = (n-1)!$ for $n \in \mathbb{Z}_\geq$). We will not derive this here, but suffice it to say that it can be derived from (80) and characteristic functions.

The χ^2 distribution is useful in hypothesis testing. If we hypothesize that a sequence of random variables X_j are normally distributed, by definition their χ^2 statistic should fall in this distribution. We can use (81) to compute the cumulative probability of this the value obtained, which we expect to be roughly k. If the value we obtain for Q is much smaller than k, it is an indication that the X_j have a more sharply peaked distribution than we expected (perhaps indicating that our predictions were "suspiciously good"). If the value we obtain is much greater than k, it indicates that our predictions are poorer than expected, perhaps ruling out our fit.

4.2.2 General Γ Distribution

Name	Parameters	Support	Mean	Variance
$\Gamma(\alpha, \beta)$	$\alpha > 0, \beta > 0$	$\mathbb{R}_\geq$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$

We have seen that the χ^2 distribution is of the form $x^a e^{-bx}$. The class of distributions of this form, with proper normalization are Γ distributions, with the general form

$$p(x) = \frac{x^{\alpha-1} e^{-\beta x} \beta^\alpha}{\Gamma(\alpha)} \qquad \forall x \geq 0 \tag{82}$$

Note that we are somewhat confusingly re-using the symbol Γ for both the Γ function and the Γ distribution.

4.3 B (Beta) Distribution

Name	Parameters	Support	Mean	Variance
$\mathrm{B}(\alpha, \beta)$	$\alpha > 0, \beta > 0$	$[0, 1]$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

I will mention the B distribution only briefly (this should be read as an uppercase β, not the latin capital "B"). This distribution is defined only on $[0, 1]$, and is given by

$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad \forall x \in [0, 1] \tag{83}$$

This distribution is notable mostly because it is a class of analytic distributions on $[0, 1]$. It is frequently used as a Bayesian prior distribution for parameters on finite intervals.

4.4 Bernoulli Distribution

Name	Parameters	Support	Mean	Variance
$\mathrm{Bern}(p)$ (not standard)	$p \in [0, 1]$	$\mathbb{Z}_2$	p	$p(1-p)$

We have yet to discuss the simplest possible non-trivial probability sample set: one with $|\Omega| = 2$. Due to the properties of probability, the only possible probability function over this sample set is

$$P(k) = \begin{cases} p & \Leftarrow k = 1 \\ 1-p & \Leftarrow k = 0 \end{cases} \tag{84}$$

We have expressed the argument k as an integer, but of course we can define this distribution for any cardinality 2 set by mapping ot to $\mathbb{Z}_2$. Often it is more convenient to write this as

$$P(k) = p^k(1-p)^{1-k} \tag{85}$$

It is trivial to show that the mean of this distribution is p and its variance is $p(1-p)$ by direct calculation.

The extremely, even maximally simple form of distribution means there is not much to say about it in and of itself. However, as we will see, it can be used to construct many other distributions. In this note we will write random variables in this distribution $X \sim \text{Bern}(p)$, though we should emphasize that this is not a standard notation. Sometimes the Bernoulli distribution is written $B(p)$, but this risks confusion with the binomial or beta distributions.

4.5 Binomial Distribution

Name	Parameters	Support	Mean	Variance
$\text{Binom}(n,p)$	$n \in \mathbb{Z}_{\geq}, p \in [0,1]$	$\mathbb{Z}_{\geq}$	np	$np(1-p)$

Given a set of variables $X_j \sim \text{Bern}(p)$, the distribution of the sum

$$X = \sum_{j=1}^{n} X_j \sim \text{Binom}(n,p) \tag{86}$$

is distributed according to the binomial distribution, given by

$$P(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{87}$$

The coefficient $\frac{n!}{k!(n-k)!}$ is known as the binomial coefficient, often written $\binom{n}{k}$. More prosaically, the binomial distribution is the distribution of the number of "successes" (results of 1) in n samples of the Bernoulli distribution with probability p (which can also be thought of as "weighted coin flips"). Since X is expressible as a sum over i.i.d. variables, in accordance with the CLT, it approaches as Gaussian with mean np and variance $np(1-p)$ for large n, meaning that the binomial distribution approaches the Gaussian distribution for large n.

4.6 Poisson Distribution

Name	Parameters	Support	Mean	Variance
$\text{Pois}(\lambda)$	$\lambda > 0$	$\mathbb{Z}_{\geq}$	λ	λ

The Poisson distribution is that of numbers of events which occur with a "fixed probability per unit time", a notion we will make more precise in the following.

Suppose that we sample a Bernoulli distribution once every time interval of duration δt. Let Δ be a fixed time duration, and let $n\,\delta t = \Delta$. We can take the limit $\delta t \to 0$, or equivalently $n \to \infty$ while Δ is held fixed. Then let $p = \frac{\lambda}{n}$ for each Bernoulli distribution each δt interval. As we have seen, this is merely a binomial distribution with $p = \frac{\lambda}{n}$, so that the probability function is

$$\begin{aligned} P(k) &= \lim_{n\to\infty} \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \lim_{n\to\infty} \frac{n^k + \mathcal{O}\left(n^{k-1}\right)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \lim_{n\to\infty} \frac{\lambda^k}{k!} \left(1-\frac{\lambda}{n}\right)^{n-k} \end{aligned} \tag{88}$$

In the first line we have simply written the binomial coefficient $\frac{n!}{k!(n-k)!}$ in a more suggestive way. Using the identity $\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^k = e^{-\lambda}$ we find a simple expression for the Poisson distribution

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{89}$$

This makes more rigorous our whimsical notion of "fixed probability per unit time". We could have written the derivation above in terms of $(\frac{\lambda}{\Delta})\delta t$, taking the limit $\delta t \to 0$ (equivalent to $n \to \infty$). In this sense $\frac{\lambda}{\Delta}$ can be interpreted as a probability per unit time. Note that the parameter $\lambda > 0$ is not bounded above, as we may have many occurrences of an event in time Δ.

We might expect that, since we have derived it by taking the $n \to \infty$ limit of the binomial distribution, that the Poisson distribution is merely the Gaussian distribution. The reason this is not the case is essentially that the domain of the Poisson distribution is $\mathbb{Z}_{\geq}$, not $\mathbb{R}$ as for the Gaussian (i.e. not only integer but also *positive*). Indeed, the CLT manifests itself in that the Poisson distribution approaches the Guassian distribution for large λ.

5 Estimators and Hypothesis Testing

We cannot observe probability distributions directly, instead they must be inferred form statistical properties of samples. For example, if by hypothesis we observe the outcome of a process that's distributed according to $\mathcal{N}(\mu, \sigma)$, we cannot determine μ or σ exactly, but instead must estimate them from the properties of a finite sample. A statistic that we use to estimate a distribution parameter is called an **estimator**. These are of great practical importance, so we will discuss them in more detail in this section.

5.1 Expected Error

As estimators are computed from finite samples, they will themselves have statistical properties that must be understood for them to be useful. An estimator can be expressed as some function of random variables $\hat{\theta}(X_1, ..., X_n)$. Estimators are frequently denoted with a $\hat{}$, a convention we will adhere to here. We may hide the X_j dependence for convenience where there is no risk of confusion. We'd like to quantify the difference between our estimator $\hat{\theta}$ and some parameter of the hypothesized distribution θ, so it makes sense to consider the **mean quared error** (MSE) defined as $\left\langle \left(\hat{\theta} - \theta\right)^2 \right\rangle$ (note that this is *only* the same thing as the variance of θ is indeed the mean of $\hat{\theta}$, a topic we will broach momentarily). Note that

$$\mathrm{MSE}\left[\hat{\theta}\right] \coloneqq \left\langle \left(\hat{\theta} - \theta\right)^2 \right\rangle = \left\langle \hat{\theta}^2 \right\rangle - 2\theta \left\langle \hat{\theta} \right\rangle + \theta^2 \tag{90}$$

but by splitting the first term on the right hand side using $\mathrm{var}\left[\hat{\theta}\right] = \left\langle \hat{\theta}^2 \right\rangle - \left\langle \hat{\theta} \right\rangle^2$ and re-combining terms we find

$$\mathrm{MSE}\left[\hat{\theta}\right] = \left(\left\langle \hat{\theta} \right\rangle - \theta\right)^2 + \mathrm{var}\left[\hat{\theta}\right] \tag{91}$$

The square root of the first term $\left\langle \hat{\theta} \right\rangle - \theta$ is called the **bias**, and is simply the difference between the mena of the estimator and the parameter it is supposed to estimate. (91) shows an important fundamental limitation of all statistical estimators. Both the bias and the variance contribute to the expected error. Usually we seek an estimator which minimizes the MSE. In many cases, we may use an **unbiased** estimator, meaning an estimator for which $\left\langle \hat{\theta} \right\rangle = \theta$, however, in practice this often comes at the expense of $\mathrm{var}\left[\hat{\theta}\right]$. This is an important concept in statistical learning known as the **bias-variance tradeoff**. As we will discuss in more detail in a future section in empirical risk minimization, typically more "finely detailed" models with a larger number of parameters have a small bias but large variance, whereas simpler models with fewer degrees of freedom have smaller variance but in some cases may have irreducible bias.

5.2 Sample Mean

As we have discussed, the CLT is suggestive of a way to estimate the mean $\langle X \rangle$. Since the distribution of $\sum_j X_j$ for i.i.d. random variables approaches $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, we can take as an estimator of the sample mean $(X_1 + \cdots + X_n)/n$. That is, given a series of random variables $X_1, ..., X_n$, we take as our sample mean estimator

$$\hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} X_j \tag{92}$$

and the CLT guarantees that $\langle \hat{\mu} \rangle = \mu$ where μ is the true mean of the population distribution. If the population distribution is Gaussian, also by the CLT we have $\mathrm{var}[\hat{\mu}] = \frac{\sigma^2}{n}$, otherwise this will be approximately true for large n.

Note that in the above it may appear that we conflated random variables X_j with "observation measurements". We can do this becuase the ensemble of all sequences of measurements is the same as the statistical ensemble of the random variables $X_1, ..., X_n$, under the assumption that observations are i.i.d. In other words, we should

express estimators as a function of random variables, a particular sequence of measurements corresponds to a particular realization of the sequence of random variables.

5.3 Sample Variance

From the definition of variance, we might assume a good estimator might be

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}\left(X_j - \hat{\mu}\right)^2 \tag{93}$$

(the use of a $\sim$ rather than a $\hat{}$ here is deliberate). The mean of this estimator is

$$\begin{aligned}\langle\tilde{\sigma}^2\rangle &= \frac{1}{n}\sum_{j=1}^{n}\left(\langle X_j^2\rangle - 2\langle X_j\hat{\mu}\rangle + \langle\hat{\mu}^2\rangle\right)\\ &= \frac{1}{n}\sum_{j=1}^{n}\left(\mu^2 + \sigma^2 - 2\langle X_j\hat{\mu}\rangle + \mu^2 + \frac{\sigma^2}{n}\right)\\ &= 2\mu^2 + \frac{n+1}{n}\sigma^2 - 2\langle X\hat{\mu}\rangle\end{aligned} \tag{94}$$

where $\mu = \langle X\rangle$ and $\sigma^2 = \mathrm{var}[X]$. Here we use X to mean any X_j since they are identically distributed. We have also repeatedly used $\langle X^2\rangle = \mu^2 + \sigma^2$. Now we need

$$\begin{aligned}\langle X_j\hat{\mu}\rangle = \frac{1}{n}\sum_{k=1}^{n}\langle X_jX_k\rangle &= \frac{1}{n}\left(\langle X^2\rangle + \sum_{k\neq j}\langle X_jX_k\rangle\right)\\ &= \frac{1}{n}(\mu^2 + \sigma^2 + (n-1)\mu^2)\\ &= \mu^2 + \frac{\sigma^2}{n}\end{aligned} \tag{95}$$

Combining this with (94) we find

$$\langle\tilde{\sigma}^2\rangle = 2\mu^2 + \frac{n+1}{n}\sigma^2 - 2\mu^2 - \frac{2\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \tag{96}$$

From this we see that $\tilde{\sigma}^2$ is biased in that its mean differs from σ^2 by a factor of $(n-1)/n$. Knowing this, it is trivial to construct the unbiased estimator

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \hat{\mu}\right)^2 \tag{97}$$

This result is known as the scourge of statistics undergraduates everywhere. Fortunately, for large n the bias of $\tilde{\sigma}^2$ becomes negligible, so the discrepancy between this and the unbiased estimator $\hat{\sigma}^2$ tends not to be relevant in applications with appreciable sample sizes. Note that $\hat{\mu}$, not μ appears in our expression for $\hat{\sigma}^2$, since we can oberve $\hat{\mu}$ but not μ.

As one should expect, $\mathrm{var}[\hat{\sigma}^2] \propto \sigma^4$. Deriving the estimator variance is tedious, but straightforward, so we merely state it here without proof

$$\mathrm{var}[\hat{\sigma}^2] = \frac{\sigma^4}{n}\left(m_4 - 1 + \frac{2}{n-1}\right) \tag{98}$$

where m_4 is the fourth moment of the population distribution. Notably, for a Gaussian population distribution this reduces to $\mathrm{var}[\hat{\sigma}^2] = \frac{2\sigma^4}{n-1}$. In either case, it is useful to know that the variance of this estimator decreases as $1/n$, which is the same as for $\mathrm{var}[\hat{\mu}]$.

5.4 Maximum Likelihood Estimation (MLE)

To fit a theory with parameters θ to observation data x, we should maximize the probability $P(x|\theta)$. This procedure is known as a **likelihood fit**, or **maximum likelihood estimation**. Usually, rather than maximizing $P(x|\theta)$ directly, we maximize its logarithm, and define the **log likelihood function**

$$\ell(\theta) = \log[P(x|\theta)] \tag{99}$$

The maximum likelihood estimate is then

$$\hat{\theta} = \arg\max_{\theta}\ell(\theta) \tag{100}$$

The logarithm is useful for numerical stability, and because many probability distributions are either exponential, can be easily factorized, or both.

5.4.1 With Gaussian Noise

Consider, as a hypothesis

$$Y = f(X;\theta) + \varepsilon \tag{101}$$

where X, Y and ε are random variables. By assumption $\varepsilon \sim \mathcal{N}(0,\sigma)$ is independent of both X and Y. θ is a set of model parameters which we wish to estimate by maximizing the probability of a sequence of observations $O = \{(x_1,y_2),(x_2,y_2),...,(x_n,y_n)\}$.

Since $Y - f(X;\theta) = \varepsilon$, the random variable $Y - f(X;\theta) \sim \mathcal{N}(0,\sigma)$. By assumption, each observation (x,y) is a sample of (X,Y). Then

$$P(O|\theta) \propto \prod_{j=1}^{n} \exp\left[-\frac{1}{2}\left(\frac{y_j - f(x_j;\theta)}{\sigma}\right)^2\right] \tag{102}$$

We will not bother computing the normalization factor for $P(O|\theta)$ since our goal will be to minimize $\ell(\theta)$, which we will define without regard to normalization

$$\ell(\theta) = \sum_{j=1}^{n} \left(\frac{y_j - f(x_j;\theta)}{\sigma}\right)^2 \tag{103}$$

This tells us that the maximum likelihood model can be determined by minimizing the sum of squares of the model error for all observations $y_j - f(x_j;\theta)$, weighted by the variance σ. In realistic cases, the assumed error variance σ^2 is not necessarily the same for each observation. We can generalize to this case by assuming $Y_j = f(X_j;\theta) + \varepsilon_j$ where the X_j's and Y_j's are by hypothesis each distributed the same as X and Y respectively, but the $\varepsilon_j \sim \mathcal{N}(0,\sigma_j)$. Therefore, we can generalize (103) to

$$\ell(\theta) = \sum_{j=1}^{n} \left(\frac{y_j - f(x_j;\theta)}{\sigma_j}\right)^2 \tag{104}$$

By minimizing ℓ with respect to θ we now have a general formual for fitting model parameters θ to observations. This required us to assume the distribution of ε, but notably it did not require us to assume any particular form of f. This formula is therefore valid for any hypothesized f. In the case where $f(x;\theta)$ is linear in both x and θ, the model which minimizes $\ell(\theta)$ is known as a **linear regression**.

5.4.2 Of a Binary Variable

Now we suppose instead that $Y : \Omega \to \mathbb{Z}_2$ and $X : \Omega \to \mathbb{R}$. We have already seen that the unique probability distribution for a sample set with cardinality 2 is the Bernoulli distribution, which has a single parameter p. We take as our hypothesis

$$Y \sim \text{Bern}(p) \qquad p = f(X;\theta) \tag{105}$$

so that

$$P(O|\theta) \propto \prod_{j=1}^{n} f^{y_j}(x_j;\theta)\big(1 - f(x_j;\theta)\big)^{1-y_j} \tag{106}$$

Then, taking the log of the Bernoulli distribution, up to a constant we have

$$\ell(\theta) = \sum_{j=1}^{n} \big[y_j \log\big(f(x_j;\theta)\big) + \big(1-y_j\big)\log\big(1 - f(x_j;\theta)\big)\big] \tag{107}$$

For this to make sense, we must have $0 \le f(x;\theta) \le 1$. One possible choice for f is

$$f(x;\theta) = \frac{1}{1 + e^{-h(x;\theta)}} \tag{108}$$

where $h : \mathbb{R} \to \mathbb{R}$. The function $(1+e^{-x})^{-1}$ is known as the **logit** function. The model which minimizes ℓ in the case where $h(x;\theta)$ is linear in both x and θ is called a **logistic regression**. Note that

$$h(x;\theta) = \log\left[\frac{f(x;\theta)}{1 - f(x;\theta)}\right] \tag{109}$$

Since f is playing the role of a probability (the parameter p of the Bernoulli distribution), h is also known as the **log odds**.

Our choice of f was of course completely arbitrary. There are other common choices, for example

$$f(x;\theta) = \frac{1}{2}[1 + \text{erf}(h(x;\theta))] \tag{110}$$

where h is linear in both x and θ (in other words, f is the cumulative probability distribution of a Gaussian). The CLT serves as motivation for this choice. This is called a **probit regression**. $f(x;\theta)$ and Φ are qualitatively similar, but the former is somewhat easier to deal with. The similarity seems notable because the logistic regression is far more common than the probit regression, but only in light of the latter does the motivation for the former become apparent.

5.4.3 Generalized Linear Models

Generalizing the previous two examples, we can write

$$\langle Y \rangle_\varepsilon = g^{-1}(\beta X) \tag{111}$$

The expectation value on the left hand side is to be taken over stochastic noise assumed to connect the model to observations, which here and in the Gaussian example we denote ε. X can have any number of dimensions and β is a linear operator. The function $g : \mathbb{R} \to \mathbb{R}$ is called a **link function**. We can think of this as a function of a "mean value" given in terms of X, for example if $g = \mathbf{1}$, the expectation value of Y is simply βX as in our first example.

Technically, the distribution of ε over which we take $\langle Y \rangle$ needn't depend on the link function g. However, the range of g can inform our choice. In our first example, $\text{range}(g) = \mathbb{R}$, so we chose $\varepsilon \sim \mathcal{N}(0, \sigma)$. In the logistic regression, the $\text{range}(g) = [0, 1]$ so we chose a Bernoulli distribution, though we could just as easily have chosen the beta distribution. Some common pairings of distribution and link function are shown in Table 9.

Distribution	**Support**	**Link Function** $g(\mu)$	**Mean** $\langle Y \rangle$
$\mathcal{N}(\mu, \sigma^2)$	$\mathbb{R}$	1	βX
$\Gamma(\alpha, \beta)$	$\mathbb{R}_{\geq}$	$-\frac{1}{\mu}$	$-(\beta X)^{-1}$
$\text{Pois}(\lambda)$	$\mathbb{Z}_{\geq}$	$\log(\mu)$	$e^{\beta X}$
$\text{Bern}(p)$	$\mathbb{Z}_2$	$\log\left(\frac{\mu}{1-\mu}\right)$	$\left(1 + e^{-\beta X}\right)^{-1}$
$\text{Binom}(n, p)$	$\mathbb{Z}_n$	$\log\left(\frac{\mu}{n-\mu}\right)$	$\left(1 + e^{-\beta X}\right)^{-1}$

Table 9: Common choices of link function.

5.4.4 Empirical Risk Minimization (ERM)

In the section on maximum likelihood estimation we have repeatedly, without explicit justification, split the distribution of observed data into the variables X and Y. While technically not a requirement, the reason for doing this is that experiments typically have some controlled variable represented by X and some response variable represented by Y. These must fall in some joint probability distribution $p_{XY}(x, y)$. Another way of finding a model of best fit is to define some **loss function** L and minimize its expectation, for example

$$\langle L(Y, f(X;\theta)) \rangle = \int \mathrm{d}P_{XY}(x, y)\, L(y, h(x;\theta)) \tag{112}$$

Typically L is a function that characterizes the "badness" of a model described by the function $f(\cdot;\theta)$, for example the mean squared error. Typically the integral is estimated by summing over observed (x, y) pairs and the model parameters θ are determined by minimizing $\langle L \rangle$.

5.5 Monte Carlo Integration

Consider the probability distribution

$$p(x) = \begin{cases} \frac{1}{\text{vol}(V)} & \Leftarrow x \in V \\ 0 & \Leftarrow x \notin V \end{cases} \tag{113}$$

for $x \in \mathbb{R}^n$ with $V \subset \mathbb{R}^n$ and $\text{vol}(V)$. Then, for any function $f : \mathbb{R} \to \mathbb{R}^n$

$$\langle f(X) \rangle = \int_V \mathrm{d}^n x\, f(x) \tag{114}$$

We can exploit this relationship to compute the integral $\int_V \mathrm{d}^n x\, f(x)$ using an estimator for the mean of f with respect to the random variable X. As we have already seen, one such estimator is

$$\hat{f} = \frac{1}{m}\sum_{j=1}^{m} f(x_j) \tag{115}$$

As we have seen $\mathrm{var}\left[\hat{f}\right] \propto m$, so that the statistical uncertainty in our integral decreases as $1/m$.

While this method of integration may seem very inefficient, note that the error scales with $1/m$ *regardless of the number of dimensions* n. This method, called **monte carlo integration**, is therefore very efficient in large dimensions.

Samples drawn where $f(x)$ is small must have a small contribution to $\langle f(X)\rangle$, so we'd like to avoid sampling too much in those regions. We can improve the method by exchanging (113) for something else. We do this by, instead of evaluating $\langle f(X)\rangle$, evaluating

$$\left\langle \frac{f(X')}{p_{X'}(X')} \right\rangle = \int_V \mathrm{d}^n x\, p_{X'}(x)\left(\frac{f(x)}{p_{X'}(x)}\right) = \int_V \mathrm{d}^n x\, f(x) \tag{116}$$

While the left hand side looks peculiar because of the presence of $p_{X'}$ inside the expectation value, it is merely a function, so we are free to compute whatever expectation value we like. From this we see that an improved estimator is

$$\hat{f}' = \frac{1}{m}\sum_{j=1}^{m} \frac{f(x_j)}{p_{X'}(x_j)} \tag{117}$$

for which $\mathrm{var}\left[\hat{f}'\right] = \mathrm{var}[f(X')/p_{X'}(X')]/m$. This of course scales with $1/m$ just as $\mathrm{var}\left[\hat{f}\right]$ does, but we can now try to reduce the variance by choosing a $p_{X'}$ which maintains as close to constant ration with f as possible. One might be tempted to then simply pick $p_{X'}(x) \propto f(x)$, but to do this we'd have to normalize $f(x)$, which requires computing its integral, which is what we're trying to do in the first place. Despite this, way may still be able to significanltly reduce the variance of the estimator by choosing a $p_{X'}$ which can be efficiently sampled from.

6 Stochastic Processes

A stochastic processes is an uncountable set of random variables parameterized by some parameter t. For example $\{X(t) \mid \forall t \in \mathbb{R}^n\}$ where each $X(t)$ is a random variable. That is, we can think of X as $X(t;\omega)$ where $\omega \in \Omega$ for a sample space Ω such that $X(t;\cdot) : \Omega \to \mathcal{M}$ is a random variable for each $t \in \mathbb{R}^n$. Another way to think about this is as a "function-valued random variable". Obviously, this is a very broad class of objects samples of which can include both smooth and discontinuous functions. To characterize them, we must consider the probability density of each $X(t)$. For notational convenience we define

$$m_X(t) = \langle X(t)\rangle \tag{118}$$

In general, the probability densities at different points are related: that is $X(t_1)$ and $X(t_2)$ can be correlated. Therefore, the covariance

$$K_X(t_1,t_2) = \mathrm{cov}[X(t_1), X(t_2)] \tag{119}$$

plays a key role in the study of stochastic processes. We refer to K_X as the *covariance function* for the random process X. In the future we will drop the subscript X where convenient and where there is no risk of confusion.

A process is called *strictly stationary* if its distribution does not depend on t. More precisely, this means that the joint distribution of $\{X(t) \mid t \in A\}$ is equal to the joint distribution of $\{X(t) \mid t \in B\}$ $\forall A, B \in \mathbb{R}^n$. A weaker notion of stationarity, sometimes called *wide stationarity* requires that

$$m(t) = \text{const} \qquad \wedge \qquad K(t_1,t_2) = K(g(t_1,t_2)) \tag{120}$$

where $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a metric. Of course, K can also possess additional symmetries, such as rotation, and we can classify Gaussian processes by the symmetry properties of K.

6.1 Gaussian Processes

An important special case of a stochastic process is a *Gaussian process* in which $X(t) \sim \mathcal{N}_n(m(t), K(t,t))$ $\forall t$. In this case, $m(t)$ and $K(t_1,t_2)$ fully specify the distribution of the process X because these are the only parameters of the distribution.

The properties of multivariate Gaussians that we discussed in Section 4.1.2 now tell us everything we need to know about the Gaussian process X. As an important example of this, suppose we have observed $X(t)$ at some set of parameters $t_1, t_2, ..., t_l$. To compute the conditional distribution $X(t) \mid X(t_1), X(t_2), \cdots, X(t_l)$ we define

$$\begin{gathered} \Sigma = \begin{pmatrix} K(t,t) & \Sigma^*(t) \\ \Sigma^{*\mathrm{T}}(t) & \Sigma^{**} \end{pmatrix} \\ \Sigma_j^* = K\big(t, t_j\big) \qquad \Sigma_{ij}^{**} = K\big(t_i, t_j\big) \qquad (1 \le i \le l)(1 \le j \le l) \end{gathered} \tag{121}$$

These can be simply plugged into (79) to obtain the conditional distribution explicitly.

Strict stationarity and wide stationarity, discussed above, are equivalent for Gaussian processes.