1

Mathematical Statistics and Information Theory

There are many easily found good books on probability theory and mathematical statistics (eg [84, 85, 87, 117, 120, 122, 196]), stochastic processes (eg [31, 161]) and information theory (eg [175, 176]); here we just outline some topics to help make the sequel more self contained. For those who have access to the computer algebra package *Mathematica* [215], the approach to mathematical statistics and accompanying software in Rose and Smith [177] will be particularly helpful.

The word stochastic comes from the Greek *stochastikos*, meaning skillful in aiming and *stochazesthai* to aim at or guess at, and *stochos* means target or aim. In our context, stochastic colloquially means involving chance variations around some event—rather like the variation in positions of strikes aimed at a target. In its turn, the later word statistics comes through eighteenth century German from the Latin root *status* meaning state; originally it meant the study of political facts and figures. The noun random was used in the sixteenth century to mean a haphazard course, from the Germanic randir to run, and as an adjective to mean without a definite aim, rule or method, the opposite of purposive. From the middle of the last century, the concept of a random variable has been used to describe a variable that is a function of the result of a well-defined statistical experiment in which each possible outcome has a definite probability of occurrence. The organization of probabilities of outcomes is achieved by means of a probability function for discrete random variables and by means of a probability density function for continuous random variables. The result of throwing two fair dice and summing what they show is a discrete random variable.

Mainly, we are concerned with continuous random variables (here measurable functions defined on some $\mathbb{R}^n$) with smoothly differentiable probability density measure functions, but we do need also to mention the Poisson distribution for the discrete case. However, since the Poisson is a limiting approximation to the Binomial distribution which arises from the Bernoulli distribution (which everyone encountered in school!) we mention also those examples.

K. Arwini, C.T.J. Dodson, *Information Geometry.*
Lecture Notes in Mathematics 1953,

1.1 Probability Functions for Discrete Variables

For discrete random variables we take the domain set to be $\mathbb{N} \cup \{0\}$. We may view a probability function as a subadditive measure function of unit weight on $\mathbb{N} \cup \{0\}$

$$p \; : \; \mathbb{N} \cup \{0\} \to [0,1) \qquad \text{(nonnegativity)} \tag{1.1}$$

$$\sum_{k=0}^{\infty} p(k) = 1 \qquad \text{(unit weight)} \tag{1.2}$$

$$p(A \cup B) \leq p(A) + p(B), \; \forall A, B \subset \mathbb{N} \cup \{0\}, \qquad \text{(subadditivity)} \tag{1.3}$$

$$\text{with equality} \iff A \cap B = \emptyset.$$

Formally, we have a discrete measure space of total measure 1 with σ-algebra the power set and measure function induced by p

$$sub(\mathbb{N} \cup \{0\}) \to [0,1) : A \mapsto \sum_{k \in A} p(k)$$

and as we have anticipated above, we usually abbreviate $\sum_{k \in A} p(k) = p(A)$.

We have the following expected values of the random variable and its square

$$\mathcal{E}(k) = \overline{k} = \sum_{k=0}^{\infty} k \, p(k) \tag{1.4}$$

$$\mathcal{E}(k^2) = \overline{k^2} = \sum_{k=0}^{\infty} k^2 \, p(k). \tag{1.5}$$

Formally, statisticians are careful to distinguish between a property of the whole population—such as these expected values—and the observed values of samples from the population. In practical applications it is quite common to use the bar notation for expectations and we shall be clear when we are handling sample quantities. With slight but common abuse of notation, we call $\overline{k}$ the mean, $\overline{k^2} - (\overline{k})^2$ the variance, $\sigma_k = +\sqrt{\overline{k^2} - (\overline{k})^2}$ the standard deviation and $\sigma_k / \overline{k}$ the coefficient of variation, respectively, of the random variable k. The variance is the square of the standard deviation.

The moment generating function $\Psi(t) = \mathcal{E}(e^{tX}), \; t \in \mathbb{R}$ of a distribution generates the r^{th} moment as the value of the r^{th} derivative of Ψ evaluated at $t = 0$. Hence, in particular, the mean and variance are given by:

$$\mathcal{E}(X) = \Psi'(0) \tag{1.6}$$

$$Var(X) = \Psi''(0) - (\Psi'(0))^2, \tag{1.7}$$

which can provide an easier method for their computation in some cases.

1.1.1 Bernoulli Distribution

It is said that a random variable X has a Bernoulli distribution with parameter p $(0 \leq p \leq 1)$ if X can take only the values 0 and 1 and the probabilities are

$$P_r(X=1) = p \tag{1.8}$$

$$P_r(X=0) = 1-p \tag{1.9}$$

Then the probability function of X can be written as follows:

$$f(x|p) = \begin{cases} p^x(1-p)^{1-x} & \text{if } x = 0,1 \\ 0 & \text{otherwise} \end{cases} \tag{1.10}$$

If X has a Bernoulli distribution with parameter p, then we can find its expectation or mean value $\mathcal{E}(X)$ and variance $Var(X)$ as follows.

$$\mathcal{E}(X) = 1 \cdot p + 0 \cdot (1-p) = p \tag{1.11}$$

$$Var(X) = \mathcal{E}(X^2) - (\mathcal{E}(X))^2 = p - p^2 \tag{1.12}$$

The moment generating function of X is the expectation of e^{tX},

$$\Psi(t) = \mathcal{E}(e^{tX}) = pe^t + q \tag{1.13}$$

which is finite for all real t.

1.1.2 Binomial Distribution

If n random variables $X_1, X_2, \ldots, X_n$ are independently identically distributed, and each has a Bernoulli distribution with parameter p, then it is said that the variables $X_1, X_2, \ldots, X_n$ form n Bernoulli trials with parameter p.

If the random variables $X_1, X_2, \ldots, X_n$ form n Bernoulli trials with parameter p and if $X = X_1 + X_2 + \ldots + X_n$, then X has a binomial distribution with parameters n and p.

The binomial distribution is of fundamental importance in probability and statistics because of the following result for any experiment which can have outcome only either success or failure. The experiment is performed n times independently and the probability of the success of any given performance is p. If X denotes the total number of successes in the n performances, then X has a binomial distribution with parameters n and p. The probability function of X is:

$$P(X=r) = P(\sum_{i=1}^{n} X_i = r) = \binom{n}{r} p^r (1-p)^{n-r} \tag{1.14}$$

where $r = 0, 1, 2, \ldots, n$.

We write

$$f(r|p) = \begin{cases} \binom{n}{r} p^r (1-p)^{n-r} & \text{if r=0, 1, 2, \ldots , n} \\ 0 & \text{otherwise} \end{cases} \tag{1.15}$$

In this distribution n must be a positive integer and p must lie in the interval $0 \leq p \leq 1$. If X is represented by the sum of n Bernoulli trials, then it is easy to get its expectation, variance and moment generating function by using the properties of sums of independent random variables—cf. §1.3.

$$\mathcal{E}(X) = \sum_{i=1}^{n} \mathcal{E}(X_i) = np \tag{1.16}$$

$$Var(X) = \sum_{i=1}^{n} Var(X_i) = np(1-p) \tag{1.17}$$

$$\Psi(t) = \mathcal{E}(e^{tX}) = \prod_{i=1}^{n} \mathcal{E}(e^{tX_i}) = (pe^t + q)^n. \tag{1.18}$$

1.1.3 Poisson Distribution

The Poisson distribution is widely discussed in the statistical literature; one monograph devoted to it and its applications is Haight [102].

Take $t, \tau \in (0, \infty)$

$$p \;:\; \mathbb{N} \cup \{0\} \to [0,1) : k \mapsto \left(\frac{t}{\tau}\right)^k \frac{1}{k!}\, e^{-t/\tau} \tag{1.19}$$

$$\overline{k} = t/\tau \tag{1.20}$$

$$\sigma_k = t/\tau. \tag{1.21}$$

This probability function is used to model the number k of events in a region of measure t when the mean number of events per unit region is τ and the probability of an event occurring in a region depends only on the measure of the region, not its shape or location. Colloquially, in applications it is very common to encounter the usage of 'random' to mean the specific case of a Poisson process; formally in statistics the term random has a more general meaning: probabilistic, that is dependent on random variables. Figure 1.1 depicts a simulation of a 'random' array of 2000 line segments in a plane; the centres of the lines follow a Poisson process and the orientations of the lines follow a uniform distribution, cf. §1.2.1. So, in an intuitive sense, this is the result of the least choice, or maximum uncertainty, in the disposition of these line segments: the centre of each line segment is equally likely to fall in every region of given area and its angle of axis orientation is equally likely to fall in every interval of angles of fixed size. This kind of situation is representative

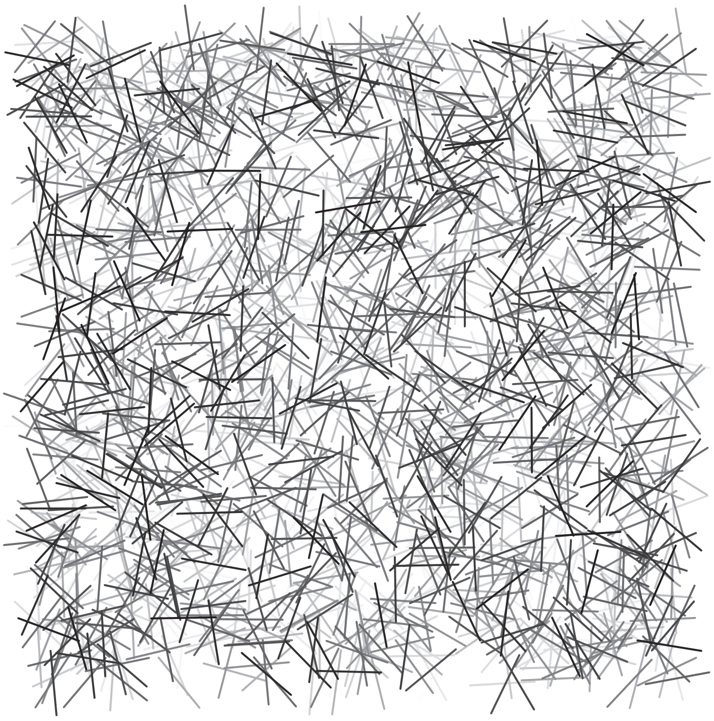

Fig. 1.1. Simulation of a random array of 2000 line segments in a plane; the centres of the lines follow a Poisson process and the orientations of the lines follow a uniform distribution. The grey tones correspond to order of deposition.

of common usage of the term 'random process' to mean subordinate to a Poisson process. A 'non-random' processes departs from Poisson by having constraints on the probabilities of placing of events or objects, typically as a result of external influence or of interactions among events or objects.

Importantly, the Poisson distribution can give a good approximation to the binomial distribution when n is large and p is close to 0. This is easy to see by making the correspondences:

$$e^{-pn} \longrightarrow (1-(n-r)p) \tag{1.22}$$

$$n!/(n-r)! \longrightarrow n^r. \tag{1.23}$$

Much of this monograph is concerned with the representation and classification of deviations from processes subordinate to a Poisson random variable, for example for a line process via the distribution of inter-event (nearest neighbour, or inter-incident) spacings. Such processes arise in statistics under the term renewal process [150].

We shall see in Chapter 9 that, for physical realisations of stochastic fibre networks, typical deviations from Poisson behaviour arise when the centres of the fibres tend to cluster, Figure 9.1, or when the orientations of their axes have preferential directions, Figure 9.15. Radiographs of real stochastic fibre networks are shown in Figure 9.3 from Oba [156]; the top network consists of fibres deposited approximately according to a Poisson planar process whereas in the lower networks the fibres have tended to cluster to differing extents.

1.2 Probability Density Functions for Continuous Variables

We are usually concerned with the case of continuous random variables defined on some $\Omega \subseteq \mathbb{R}^m$. For our present purposes we may view a probability density function (pdf) on $\Omega \subseteq \mathbb{R}^m$ as a subadditive measure function of unit weight, namely, a nonnegative map on Ω

$$f \; : \; \Omega \to [0, \infty) \qquad \text{(nonnegativity)} \tag{1.24}$$

$$\int_\Omega f = f(\Omega) = 1 \qquad \text{(unit weight)} \tag{1.25}$$

$$f(A \cup B) \leq f(A) + f(B), \; \forall A, B \subset \Omega, \qquad \text{(subadditivity)} \tag{1.26}$$
$$\text{with equality} \iff A \cap B = \emptyset.$$

Formally, we have a measure space of total measure 1 with σ-algebra typically the Borel sets or the power set and the measure function induced by f

$$sub(\Omega) \to [0,1] : A \mapsto \int_A f = \text{integral of } f \text{ over } A$$

and as we have anticipated above, we usually abbreviate $\int_A f = f(A)$. Given an integrable (ie measurable in the σ-algebra) function $u : \Omega \;\; \to \mathbb{R}$, the expectation or mean value of u is defined to be

$$\mathcal{E}(u) = \overline{u} = \int_\Omega uf.$$

We say that f is the joint pdf for the random variables $x_1, x_2, \ldots, x_m$, being the coordinates of points in Ω, or that these random variables have the joint probability distribution f. If x is one of these random variables, and in particular for the important case of a single random variable x, we have the following

$$\overline{x} = \int_\Omega xf \tag{1.27}$$

$$\overline{x^2} = \int_\Omega x^2 f. \tag{1.28}$$

Again with slight abuse of notation, we call $\overline{x}$ the mean and the variance is the mean square deviation

$$\sigma_x^2 = \overline{(x-\overline{x})^2} = \overline{x^2} - (\overline{x})^2.$$

Its square root is the standard deviation $\sigma_x = +\sqrt{\overline{x^2} - (\overline{x})^2}$ and the ratio $\sigma_x/\overline{x}$ is the coefficient of variation, of the random variable x. Some inequalities for the probability of a random variable exceeding a given value are worth mentioning.

Markov's Inequality: If x is a nonnegative random variable with probability density function f then for all $a > 0$, the probability that $x > a$ is

$$\int_a^\infty f \leq \frac{\overline{x}}{a}. \tag{1.29}$$

Chebyshev's Inequality: If x is a random variable having probability density function f with zero mean and finite variance σ^2, then for all $a > 0$, the probability that $x > a$ is

$$\int_a^\infty f \leq \frac{\sigma^2}{\sigma^2 + a^2}. \tag{1.30}$$

Bienaymé-Chebyshev's Inequality: If x is a random variable having probability density function f and u is a nonnegative non-decreasing function on $(0,\infty)$, then for all $a > 0$ the probability that $|x| > a$ is

$$1 - \int_{-a}^a f \leq \frac{\bar{u}}{u(a)}. \tag{1.31}$$

The cumulative distribution function (cdf) of a nonnegative random variable x with probability density function f is the function defined by

$$F : [0,\infty) \to [0,1] : x \mapsto \int_0^x f(t)\,dt. \tag{1.32}$$

It is easily seen that if we wish to change from random variable x with density function f to a new random variable ξ when x is given as an invertible function of ξ, then the probability density function for ξ is represented by

$$g(\xi) = f(x(\xi)) \left| \frac{dx}{d\xi} \right|. \tag{1.33}$$

If independent real random variables x and y have probability density functions f, g respectively, then the probability density function h of their sum $z = x + y$ is given by

$$h(z) = \int_{-\infty}^\infty f(x)\,g(z-x)\,dx \tag{1.34}$$

and the probability density function p of their product $r = xy$ is given by

$$p(r) = \int_{-\infty}^{\infty} f(x)\, g\left(\frac{r}{x}\right) \frac{1}{|x|} dx. \tag{1.35}$$

Usually, a probability density function depends on a set of parameters, $\theta_1, \theta_2, \ldots, \theta_n$ and we say that we have an n-dimensional family. Then the corresponding change of variables formula involves the $n \times n$ Jacobian determinant for the multiple integrals, so generalizing (1.33).

1.2.1 Uniform Distribution

This is the simplest continuous distribution, with constant probability density function for a bounded random variable:

$$u \,:\, [a,b] \to [0,\infty) : x \mapsto \frac{1}{b-a} \tag{1.36}$$

$$\overline{x} = \frac{a+b}{2} \tag{1.37}$$

$$\sigma_x = \frac{b-a}{2\sqrt{3}}. \tag{1.38}$$

The probability of an event occurring in an interval $[\alpha, \beta] \subsetneqq [a,b]$ is simply proportional to the length of the interval:

$$P(x \in [\alpha, \beta]) = \frac{\beta - \alpha}{b-a}.$$

1.2.2 Exponential Distribution

Take $\lambda \in \mathbb{R}^+$; this is called the parameter of the exponential probability density function

$$f \,:\, [0,\infty) \to [0,\infty) : [a,b] \mapsto \int_{[a,b]} \frac{1}{\lambda} e^{-x/\lambda} \tag{1.39}$$

$$\overline{x} = \lambda \tag{1.40}$$

$$\sigma_x = \lambda. \tag{1.41}$$

The parameter space of the exponential distribution is $\mathbb{R}^+$, so exponential distributions form a 1-parameter family. In the sequel we shall see that quite generally we may provide a Riemannian structure to the parameter space of a family of distributions. Sometimes we call a family of pdfs a parametric statistical model.

Observe that, in the Poisson probability function (1.19) for events on the real line, the probability of zero zero events in an interval t is

$$p(0) = e^{-t/\tau}$$

and it is not difficult to show that the probability density function for the Poisson inter-event (or inter-incident) distance t on $[0,\infty)$ is an exponential probability density function (1.39) given by

$$f : [0,\infty) \to [0,\infty) : t \mapsto \frac{1}{\tau} e^{-t/\tau}$$

where τ is the mean number of events per unit interval. Thus, the occurrence of an exponential distribution has associated with it a complementary Poisson distribution, so the exponential distribution provides for continuous variables an identifier for Poisson processes. Correspondingly, departures from an exponential distribution correspond to departures from a Poisson process. We shall see below in §1.4.1 that in rather a strict sense the gamma distribution generalises the exponential distribution.

1.2.3 Gaussian, or Normal Distribution

This has real random variable x with mean μ and variance σ^2 and the familiar bell-shaped probability density function given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \, e^{\frac{-(x-\mu)^2}{2\sigma^2}}. \tag{1.42}$$

The Gaussian distribution has the following uniqueness property: For independent random variables $x_1, x_2, \ldots, x_n$ with a common continuous probability density function f, having independence of the sample mean $\bar{x}$ and sample standard deviation S is equivalent to f being a Gaussian distribution [110].

The Central Limit Theorem states that for independent and identically distributed real random variables x_i each having mean μ and variance σ^2, the random variable

$$w = \frac{(x_1 + x_2 + \ldots + x_n) - n\mu}{\sqrt{n}\sigma} \tag{1.43}$$

tends as $n \to \infty$ to a Gaussian random variable with mean zero and unit variance.

1.3 Joint Probability Density Functions

Let f be a probability density function, defined on $\mathbb{R}^2$ (or some subset thereof). This is an important case since here we have two variables, X, Y, say, and we can extract certain features of how they interact. In particular, we define their respective mean values and their covariance, σ_{xy}:

$$\overline{x} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \; f(x,y)\, dx dy \tag{1.44}$$

$$\overline{y} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y \; f(x,y)\, dx dy \tag{1.45}$$

$$\sigma_{xy} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy \; f(x,y)\, dx dy \; - \overline{x}\;\overline{y} = \overline{xy} - \overline{x}\;\overline{y}. \tag{1.46}$$

The marginal probability density function of X is f_X, obtained by integrating f over all y,

$$f_X(x) = \int_{v=-\infty}^{\infty} f_{X,Y}(x,v)\, dv \tag{1.47}$$

and similarly the marginal probability density function of Y is

$$f_Y(y) = \int_{u=-\infty}^{\infty} f_{X,Y}(u,y)\, du \tag{1.48}$$

The jointly distributed random variables X and Y are called independent if their marginal density functions satisfy

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \quad for\, all \quad x, y \in R \tag{1.49}$$

It is easily shown that if the variables are independent then their covariance (1.46) is zero but the converse is not true. Feller [84] gives a simple counterexample: let X take values $-1, +1, -2, +2$, each with probability $\frac{1}{4}$ and let $Y = X^2$; then the covariance is zero but there is evidently a (nonlinear) dependence.

The extent of dependence between two random variables can be measured in a normalised way by means of the correlation coefficient: the ratio of the covariance to the product of marginal standard deviations:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}. \tag{1.50}$$

Note that by the Cauchy-Schwartz inequality, $-1 \leq \rho_{xy} \leq 1$, whenever it exists, the limiting values corresponding to the case of linear dependence between the variables. Intuitively, $\rho_{xy} < 0$ if y tends to increase as x decreases, and $\rho_{xy} > 0$ if x and y tend to increase together.

A change of random variables from (x, y) with density function f to say (u, v) with density function g and x, y given as invertible functions of u, v involves the Jacobian determinant:

$$g(u,v) = f(x(u,v), y(u,v)) \, \frac{\partial(x,y)}{\partial(u,v)}. \tag{1.51}$$

1.3.1 Bivariate Gaussian Distributions

The probability density function of the two-dimensional Gaussian distribution has the form:

$$f(x,y) = \frac{1}{2\,\pi\,\sqrt{\sigma_1\,\sigma_2 - {\sigma_{12}}^2}}\, e^W \tag{1.52}$$

with

$$W = -\frac{1}{2\,(\sigma_1\,\sigma_2 - {\sigma_{12}}^2)}\left(\sigma_2(x-\mu_1)^2 - 2\,\sigma_{12}\,(x-\mu_1)\,(y-\mu_2) + \sigma_1(y-\mu_2)^2\right),$$

where

$$-\infty < x_1 < x_2 < \infty, \quad -\infty < \mu_1 < \mu_2 < \infty, \quad 0 < \sigma_1, \sigma_2 < \infty.$$

This contains the five parameters $(\mu_1, \mu_2, \sigma_1, \sigma_{12}, \sigma_2) = (\xi^1, \xi^2, \xi^3, \xi^4, \xi^5) \in \Theta$. So we have a five-dimensional parameter space Θ.

1.4 Information Theory

Information theory owes its origin in the 1940s to Shannon [186], whose interest was in modelling the transfer of information stored in the form of binary on-off devices, the basic unit of information being one bit: 0 or 1. The theory provided a representation for the corruption by random electronic noise of transferred information streams, and for quantifying the effectiveness of error-correcting algorithms by the incorporation of redundancy in the transfer process. His concept of information theoretic entropy in communication theory owed its origins to thermodynamics but its effectiveness in general information systems has been far reaching. Information theory worked out by the communication theorists, and entropy in particular, were important in providing a conceptual and mathematical framework for the development of chaos theory [93]. There the need was to model the dynamics of adding small extrinsic noise to otherwise deterministic systems. In physical theory, entropy provides the uni-directional 'arrow of time' by measuring the disorder in an irreversible system [164]. Intuitively, we can see how the entropy of a state modelled by a point in a space of probability density functions would be expected to be maximized at a density function that represented as nearly as possible total disorder, colloquially, randomness.

Shannon [186] considered an information source that generates symbols from a finite set $\{x_i | i = 1, 2, \cdots n\}$ and transmits them as a stationary stochastic process. He defined the 'entropy' function for the process in terms of the probabilities $\{p_i | i = 1, 2, \cdots n\}$ for generation of the different symbols:

$$S = -\sum_{i=1}^{i=n} p_i \log(p_i). \tag{1.53}$$

This entropy (1.53) is essentially the same as that of Gibbs and Boltzmann in statistical mechanics but here it is viewed as a measure of the 'uncertainty' in the process; for example S is greater than or equal to the entropy conditioned by the knowledge of a second random variable. If the above symbols are generated mutually independently, then S is a measure of the amount of information available in the source for transmission. If the symbols in a sequence are not mutually independently generated, Shannon introduced the information 'capacity' of the transmission process as $C = \lim_{T\to\infty} \log N(T)/T$, where $N(T)$ is the maximum number of sequences of symbols that can be transmitted in time T. It follows that, for given entropy S and capacity C, the symbols can be encoded in such a way that $\frac{C}{S-\epsilon}$ symbols per second can be transmitted over the channel if $\epsilon > 0$ but not if $\epsilon < 0$. So again, we have a maximum principle from entropy.

Given a set of observed values $< g_\alpha(x) >$ for functions g_α of the random variable x, we seek a 'least prejudiced' set of probability values for x on the assumption that it can take only a finite number of values, x_i with probabilities $p_1, p_2, \cdots, p_n$ such that

$$< g_\alpha(x) > = \sum_{i=1}^{i=n} p_i\, g_\alpha(x_i) \quad \text{for } \alpha = 1, 2, \ldots, N \tag{1.54}$$

$$1 = \sum_{i=1}^{i=n} p_i. \tag{1.55}$$

Jaynes [107], a strong proponent of Shannon's approach, showed that this occurs if we choose those p_i that maximize Shannon's entropy function (1.53). In the case of a continuous random variable $x \in \mathbb{R}$ with probability density p parametrized by a finite set of parameters, the entropy becomes an integral and the maximizing principle is applied over the space of parameters, as we shall see below.

It turns out [201] that if we have no data on observed functions of x, (so the set of equations (1.54) is empty) then the maximum entropy choice gives the exponential distribution. If we have estimates of the first two moments of the distribution of x, then we obtain the (truncated) Gaussian. If we have estimates of the mean and mean logarithm of x, then the maximum entropy choice is the gamma distribution.

Jaynes [107] provided the foundation for information theoretic methods in, among other things, Bayes hypothesis testing—cf. Tribus et al. [200, 201]. For more theory, see also Slepian [190] and Roman [175, 176]. It is fair to point out that in the view of some statisticians, the applicability of the maximum entropy approach has been overstated; we mention for example the reservations of Ripley [173] in the case of statistical inference for spatial Gaussian processes.

In the sequel we shall consider the particular case of the gamma distribution for several reasons:

- the exponential distributions form a subclass of gamma distributions and exponential distributions represent Poisson inter-event distances
- the sum of n independent identical exponential random variables follows a gamma distribution

- the sum of n independent identical gamma random variables follows a gamma distribution
- lognormal distributions may be well-approximated by gamma distributions
- products of gamma distributions are well-approximated by gamma distributions
- stochastic porous media have been modelled using gamma distributions [72].

Other parametric statistical models based on different distributions may be treated in a similar way. Our particular interest in the gamma distribution and a bivariate gamma distribution stems from the fact that the exponential distribution is a special case and that corresponds to the standard model for an underlying Poisson process.

Let Θ be the parameter space of a parametric statistical model, that is an n-dimensional smooth family of probability density functions defined on some fixed event space Ω of unit measure,

$$\int_{\Omega} p_\theta = 1 \quad \text{for all } \theta \in \Theta.$$

For each sequence $X = \{X_1, X_2, \ldots, X_n\}$, of independent identically distributed observed values, the likelihood function lik_X on Θ which measures the likelihood of the sequence arising from different $p_\theta \in S$ is defined by

$$lik_X : \Theta \to [0,1] : \theta \mapsto \prod_{i=1}^{n} p_\theta(X_i).$$

Statisticians use the likelihood function, or log-likelihood its logarithm $l = \log lik$, in the evaluation of goodness of fit of statistical models. The so-called 'method of maximum likelihood', or 'maximum entropy' in Shannon's terms, is used to obtain optimal fitting of the parameters in a distribution to observed data.

1.4.1 Gamma Distribution

The family of gamma distributions is very widely used in applications with event space $\Omega = \mathbb{R}^+$. It has probability density functions given by

$$\Theta \equiv \{f(x;\gamma,\kappa) | \gamma, \kappa \in \mathbb{R}^+\}$$

so here $\Theta = \mathbb{R}^+ \times \mathbb{R}^+$ and the random variable is $x \in \Omega = \mathbb{R}^+$ with

$$f(x;\gamma,\kappa) = \left(\frac{\kappa}{\gamma}\right)^{\kappa} \frac{x^{\kappa-1}}{\Gamma(\kappa)} e^{-x\kappa/\gamma} \tag{1.56}$$

Then $\bar{x} = \gamma$ and $Var(x) = \gamma^2/\kappa$ and we see that γ controls the mean of the distribution while κ controls its variance and hence the shape. Indeed, the property that the variance is proportional to the square of the mean, §1.2, actually characterizes gamma distributions as shown recently by Hwang and Hu [106] (cf. their concluding remark).

Theorem 1.1 (Hwang and Hu [106]). *For independent positive random variables with a common probability density function f, having independence of the sample mean and the sample coefficient of variation is equivalent to f being the gamma distribution.*

The special case $\kappa = 1$ in (1.56) corresponds to the situation of the random or Poisson process along a line with mean inter-event interval γ, then the distribution of inter-event intervals is exponential. In fact, the gamma distribution has an essential generalizing property of the exponential distribution since it represents inter-event distances for generalizations of the Poisson process to a 'censored' Poisson process. Precisely, for integer $\kappa = 1, 2, \ldots,$ (1.56) models a process that is Poisson but with intermediate events removed to leave only every κ^{th}. Formally, the gamma distribution is the κ-fold convolution of the exponential distribution, called also the Pearson Type III distribution. The Chi-square distribution with $n = 2\kappa$ degrees of freedom models the distribution of a sum of squares of n independent random variables all having the Gaussian distribution with zero mean and standard deviation σ; this is a gamma distribution with mean $\gamma = n\sigma^2$ if $\kappa = 1, 2, \ldots$. Figure 1.2 shows a family of gamma distributions, all of unit mean, with $\kappa = \frac{1}{2},\ 1,\ 2,\ 5$.

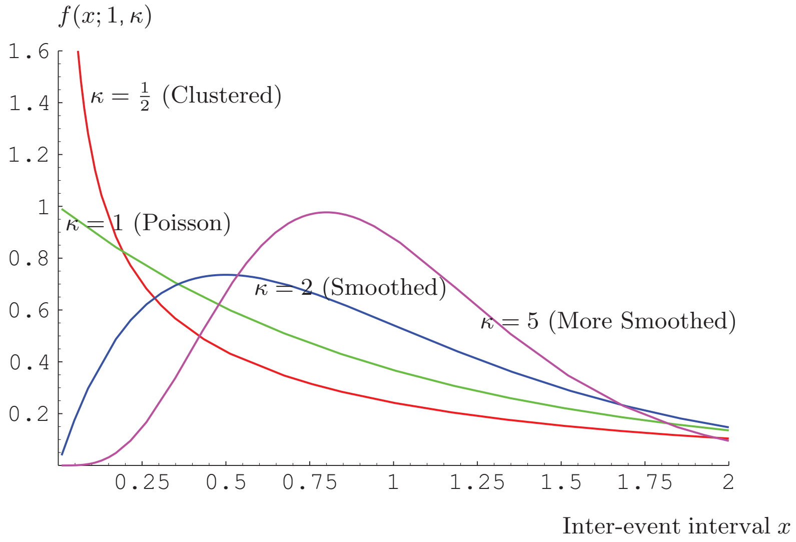

Fig. 1.2. Probability density functions, $f(x; \gamma, \kappa)$, for gamma distributions of inter-event intervals x with unit mean $\gamma = 1$, and $\kappa = \frac{1}{2},\ 1,\ 2,\ 5$. The case $\kappa = 1$ corresponds to an exponential distribution from an underlying Poisson process. Some organization—clustering ($\kappa < 1$) or smoothing ($\kappa > 1$)—is represented by $\kappa \neq 1$.

Shannon's information theoretic entropy or 'uncertainty' is given, up to a factor, by the negative of the expectation of the logarithm of the probability density function (1.56), that is

$$\begin{aligned} S_f(\gamma,\kappa) &= -\int_0^\infty \log(f(x;\gamma,\kappa))\, f(x;\gamma,\kappa)\, dx \\ &= \kappa + (1-\kappa)\frac{\Gamma'(\kappa)}{\Gamma(\kappa)} + \log\frac{\gamma\,\Gamma(\kappa)}{\kappa}. \end{aligned} \tag{1.57}$$

Part of the entropy function (1.57) is depicted in Figure 1.3 as a contour plot.

At unit mean, the maximum entropy (or maximum uncertainty) occurs at $\kappa = 1$, which is the random case, and then $S_f(\gamma, 1) = 1 + \log\gamma$. So, a Poisson process of points on a line is such that the points are as disorderly as possible and among all homogeneous point processes with a given density, the Poisson process has maximum entropy. Figure 1.4 shows a plot of $S_f(\gamma,\kappa)$, for the case of unit mean $\gamma = 1$. Figure 1.5 shows some integral curves of the entropy gradient field in the space of gamma probability density functions.

We can see the role of the log-likelihood function in the case of a set $X = \{X_1, X_2, \ldots, X_n\}$ of measurements, drawn from independent identically

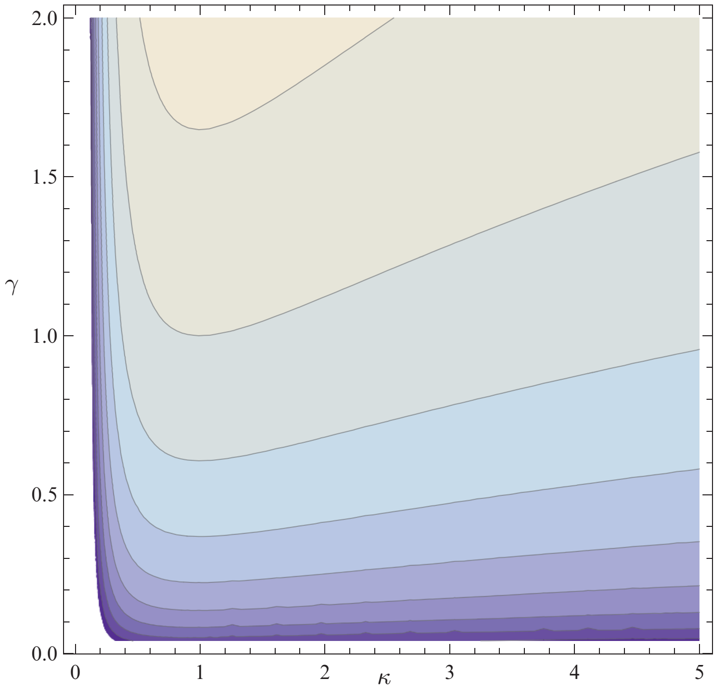

Fig. 1.3. Contour plot of information theoretic entropy $S_f(\gamma,\kappa)$, for gamma distributions from (1.57). The cases with $\kappa = 1$ correspond to exponential distributions related to underlying Poisson processes.

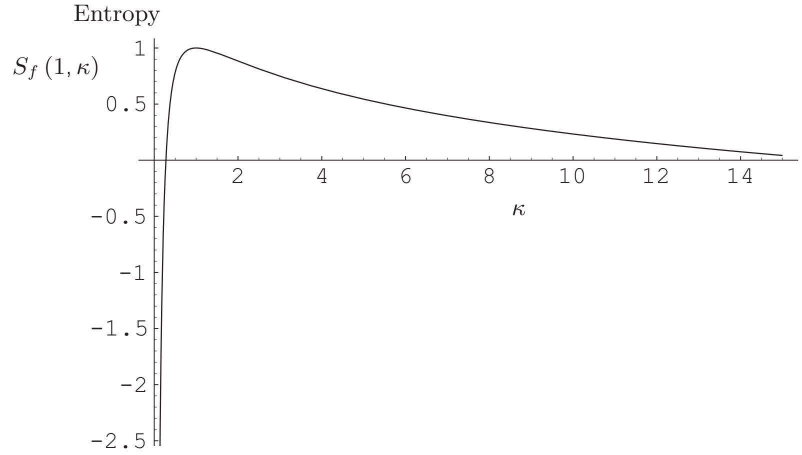

Fig. 1.4. Information theoretic entropy $S_f(\gamma,\kappa)$, for gamma distributions of inter-event intervals with unit mean $\gamma = 1$. The maximum at $\kappa = 1$ corresponds to an exponential distribution from an underlying Poisson process. The regime $\kappa < 1$ corresponds to clustering of events and $\kappa > 1$ corresponds to smoothing out of events, relative to a Poisson process. Note that, at constant mean, the variance of x decays like $1/\kappa$.

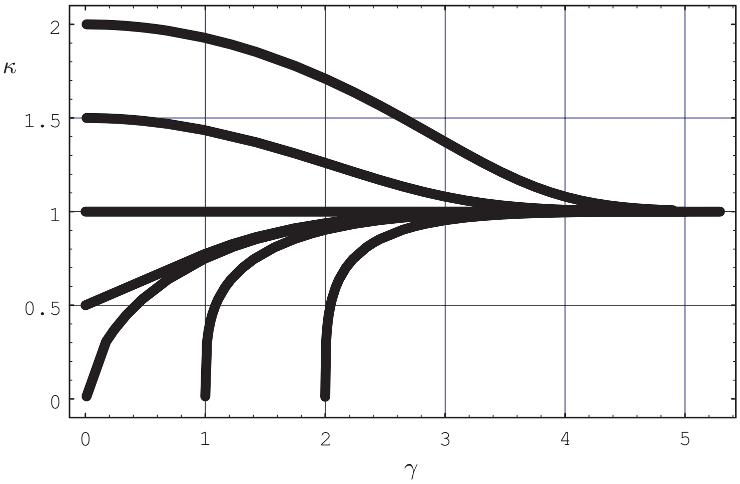

Fig. 1.5. A selection of integral curves of the entropy gradient field for gamma probability density functions, with initial points having small values of γ. The cases with $\kappa = 1$ correspond to exponential distributions related to underlying Poisson processes.

distributed random variables, to which we wish to fit the maximum likelihood gamma distribution. The procedure to optimize the choice of γ, κ is as follows. For independent events X_i, with identical distribution $p(x; \gamma, \kappa)$, their joint probability density is the product of the marginal densities so a measure of the 'likelihood' of finding such a set of events is

$$lik_X(\gamma, \kappa) = \prod_{i=1}^{n} f(X_i; \gamma, \kappa).$$

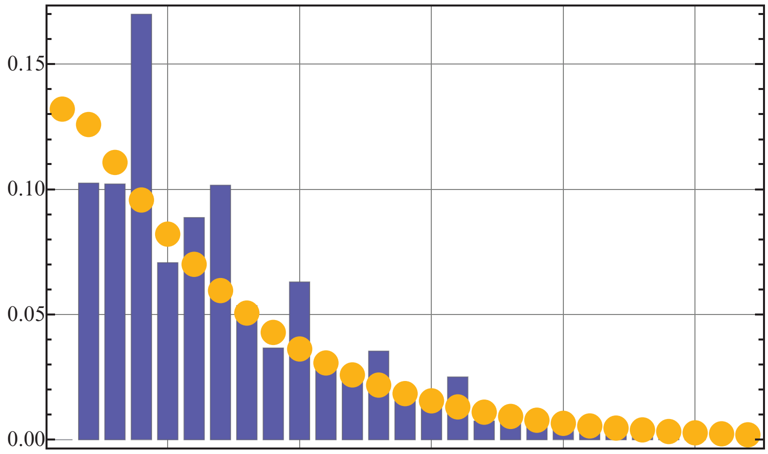

Fig. 1.6. Probability histogram plot with unit mean for the spacings between the first $100,000$ prime numbers and the maximum likelihood gamma fit, $\kappa = 1.09452$, (large points).

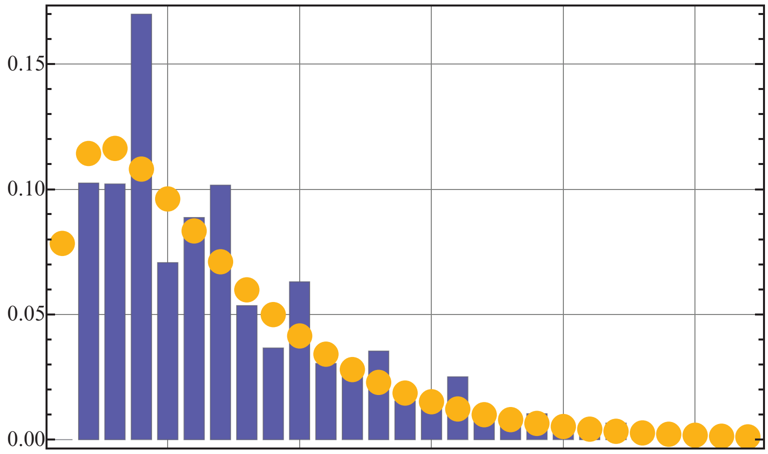

Fig. 1.7. Probability histogram plot with unit mean for the spacings between the first $100,000$ prime numbers and the gamma distribution having the same variance, so $\kappa = 1.50788$, (large points).

We seek a choice of γ, κ to maximize this product and since the log function is monotonic increasing it is simpler to maximize the logarithm

$$l_X(\gamma,\kappa) = \log lik_X(\gamma,\kappa) = \log[\prod_{i=1}^{n} f(X_i;\gamma,\kappa)].$$

Substitution gives us

$$\begin{aligned} l_X(\gamma,\kappa) &= \sum_{i=1}^{n}[\kappa(\log\kappa - \log\gamma) + (\kappa-1)\log X_i - \frac{\kappa}{\gamma}X_i - \log\Gamma(\kappa)] \\ &= n\kappa(\log\kappa - \log\gamma) + (\kappa-1)\sum_{i=1}^{n}\log X_i - \frac{\kappa}{\gamma}\sum_{i=1}^{n}X_i - n\log\Gamma(\kappa). \end{aligned}$$

Then, solving for $\partial_\gamma l_X(\gamma,\kappa) = \partial_\kappa l_X(\gamma,\kappa) = 0$ in terms of properties of the X_i, we obtain the maximum likelihood estimates $\hat{\gamma}, \hat{\kappa}$ of γ, κ in terms of the mean and mean logarithm of the X_i

$$\begin{aligned} \hat{\gamma} &= \bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i \\ \log\hat{\kappa} - \frac{\Gamma'(\hat{\kappa})}{\Gamma(\hat{\kappa})} &= \overline{\log X} - \log\bar{X} \end{aligned}$$

where $\overline{\log X} = \frac{1}{n}\sum_{i=1}^{n}\log X_i$.

For example, the frequency distribution of spacings between the first 100, 000 prime numbers has mean approximately 13.0, and variance 112, and 99% of the probability is achieved by spacings up to 4 times the mean. Figure 1.6 shows the maximum likelihood fit gamma distribution with $\kappa = 1.09452$, as points, on the probability histogram of the prime spacings normalized to unit mean; the range of the abscissa is 4 times the mean. Figure 1.7 shows as points the gamma distribution with $\kappa = 1.50788$, which has the same variance as the prime spacings normalized to unit mean. Of course, neither fit is very good and nor is the geometric distribution approximation that might be expected, cf. Schroeder [184] §4.12, in light of The Prime Number Theorem, which says that the average spacing between adjacent primes near n is approximately $\log n$.