Computational Methods in Nonlinear Physics
1. Review of probability and statistics

P.I. Hurtado

Departamento de Electromagnetismo y Física de la Materia, and Instituto Carlos I de Física Teórica y Computacional. Universidad de Granada. E-18071 Granada. Spain

E-mail: phurtado@onsager.ugr.es

Abstract. These notes correspond to the first part (20 hours) of a course on Computational Methods in Nonlinear Physics within the Master on Physics and Mathematics (*FisyMat*) of University of Granada. In this first chapter we review some concepts of probability theory and stochastic processes.

Keywords: Computational physics, probability and statistics, Monte Carlo methods, stochastic differential equations, Langevin equation, Fokker-Planck equation, molecular dynamics.

References and sources

[1] R. Toral and P. Collet, *Stochastic Numerical Methods*, Wiley (2014).
[2] H.-P. Breuer and F. Petruccione, *The Theory of Open Quantum Systems*, Oxford Univ. Press (2002).
[3] N.G. van Kampen, *Stochastic Processes in Physics and Chemistry*, Elsevier (2007).
[4] Wikipedia: https://en.wikipedia.org

Contents

1. The probability space

- This section contains a brief survey of classical probability theory and stochastic processes. Our aim is to provide a self-contained and concise presentation of the theory. The material here presented mostly follows Chapter 1 of Refs. [1] and [2].

- In most occasions, we cannot predict with absolute certainty the outcome of an experiment. This is because of (a) lack of information on initial conditions (e.g. when many degrees of freedom are involved), or (b) the underlying physics is intrinsically stochastic (as e.g. in quantum mechanics)

- Examples: number of electrons emitted by a β-radioactive substance in a given time interval, time at which a bus reaches the station, measure an electron's spin, toss a coin and look at the appearing side, or have a look through the window to observe whether it rains or not.

- In general, we have no way (or no effective way) of knowing *a priori* which one of the possible outcomes will be observed. Hence, we abandon the deterministic point of view and adopt a *probabilistic* description.

- The fundamental concept of probability theory is the probability space. It consists of three basic ingredients, namely

  – A sample space Ω of elementary events

  – A σ-algebra of events

  – A probability measure on the σ-algebra

  We shall follow here the axiomatic approach to probability which is mainly due to Kolmogorov (1956).

- Examples of sample spaces:
  - For the β-radioactive substance, $\Omega = \{0, 1, 2, ...\}$ is the set of natural numbers $\mathbb{N}$
  - The hitting times of the projectile or the arrival times of the bus (in some units) both belong to the set of real numbers $\Omega = \mathbb{R}$
  - The possible outcomes of a measure of an electron's spin are $\Omega = \{-\hbar/2, \hbar/2\}$
  - When tossing a coin, the possible results are $\Omega = \{\text{heads, tails}\}$
  - For the rain observation the set of results is $\Omega = \{\text{yes, no}\}$.

1.1. The σ-algebra of events

- We want to associate probabilities to events obtained in some kind of experiment
- Events are subsets of some basic set Ω, the sample space or space of events
- For example, if the experiment is tossing a coin, the sample space is typically the set {h(ead), t(ail)}. For tossing two coins, the corresponding sample space would be {(h,h), (h,t), (t,h), (t,t)}. An event of the σ-algebra would be, for instance, *"at least one head"*, or {(h,h), (h,t), (t,h)}$\in \Omega$. Another event would be *"no more than one head"*, i.e. {(h,t), (t,h), (t,t)}$\in \Omega$
- Another example: For tossing a single six-sided dice, the typical sample space is {1, 2, 3, 4, 5, 6} (in which the result of interest is the number of pips facing up).
- The subsets of Ω containing just one element $\omega \in \Omega$ are referred to as elementary events
- Usually we are not interested in *all* possible subsets of Ω. We rather need to specify which kind of subsets $A \subset \Omega$ we would like to include in our theory. An important requirement is that the events form a so-called σ-algebra, which is a system $\mathcal{A}$ of subsets of Ω with the following three properties:

1 The sample space itself and the empty set belong to the system of events, that is $\Omega \in \mathcal{A}$ and $\varnothing \in \mathcal{A}$.

2 If $A_1 \in \mathcal{A}$ and $A_2 \in \mathcal{A}$, then also the union $A_1 \cup A_2 \in \mathcal{A}$, the intersection $A_1 \cap A_2 \in \mathcal{A}$, and the difference $A_1 \setminus A_2 \in \mathcal{A}$ belong to the system $\mathcal{A}$.

3 If we have a countable collection of events $A_1, A_2, \ldots, A_n, \ldots \in \mathcal{A}$, then also their union $\cup_{n=1}^{\infty} A_n$ belongs to $\mathcal{A}$.

We shall always write $A \in \mathcal{A}$ to express that the subset $A \subset \Omega$ is an event of our theory.

- The above requirements ensure that the total sample space Ω and the empty set $\varnothing$ are events, and that all events of $\mathcal{A}$ can be subjected to the logical operations 'AND' ($\cap$), 'OR' ($\cup$) and 'NOT' ($\setminus$) without leaving the system of events. This is why $\mathcal{A}$ is called an algebra. The third condition is what makes $\mathcal{A}$ a σ-algebra. It tells us that any countable union of events is again an event.

- Example of σ-algebra: Flip a coin three times and count the number of heads. $\Omega = \{0, 1, 2, 3\}$. Then $\mathcal{A} = \{\varnothing, \Omega, \{0, 1, 2\}, \{3\}\}$ is a σ-algebra. Indeed, if any $A, B \in \mathcal{A}$, then $A^c \in \mathcal{A}$, $A \cup B \in \mathcal{A}$, and so on.

- Another example of σ-algebra: Flip a coin twice and count the number of heads. $\Omega = \{0, 1, 2\}$. Then $\mathcal{A} = \{\varnothing, \Omega, \{0\}, \{1\}, \{2\}, \{0, 1\}, \{0, 2\}, \{1, 2\}\}$ is a σ-algebra. Check!

1.2. Probability measures and Kolmogorov axioms

- The construction of the probability space is completed by introducing a probability measure on the σ-algebra

- A probability measure is simply a map $\mu : \mathcal{A} \to \mathbb{R}$ which assigns to each event $A \in \mathcal{A}$ of the σ-algebra a real number $\mu(A)$

$$A \to \mu(A) \in \mathbb{R} \tag{1}$$

- The number $\mu(A)$ is interpreted as the probability of the event A
- The probability measure μ is thus required to satisfy the following Kolmogorov axioms:

  1 For all events $A \in \mathcal{A}$ we have
  $$0 \leq \mu(A) \leq 1 \tag{2}$$

  2 Probability is normalized as
  $$\mu(\Omega) = 1 \tag{3}$$

  3 If we have a countable collection of disjoint events
  $$A_1, A_2, \ldots, A_n, \ldots \in \mathcal{A} \quad \text{with} \quad A_i \cap A_j = \varnothing \;\; \forall i \neq j \tag{4}$$

  then the probability of their union is equal to the sum of their probabilities,
  $$\mu(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n) \tag{5}$$

- On the basis of these axioms one can build up a consistent probability theory. In particular, the Kolmogorov axioms enable one to determine the probabilities for all events which arise from logical operations on other events. For example, one finds
  $$\mu(A_1 \cup A_2) = \mu(A_1) + \mu(A_2) - \mu(A_1 \cap A_2) \tag{6}$$

- Summary: a probability space consists of a sample space Ω, a σ-algebra $\mathcal{A}$ of events, and a probability measure μ on $\mathcal{A}$.

1.3. Conditional probabilities and statistical independence

- The concept of statistical independence is often formulated by introducing the conditional probability $\mu(A_1|A_2)$ of an event $A_1 \in \mathcal{A}$ under the condition that an event $A_2 \in \mathcal{A}$ occurred,

$$\mu(A_1|A_2) = \frac{\mu(A_1 \cap A_2)}{\mu(A_2)} \tag{7}$$

- These events are said to be statistically independent if and only if $\mu(A_1|A_2) = \mu(A_1)$, or equivalently, iff

$$\mu(A_1 \cap A_2) = \mu(A_1)\mu(A_2) \tag{8}$$

- This means that the probability of the mutual occurrence of the events A_1 and A_2 is just equal to the product of the probabilities of A_1 and A_2

- The condition of statistical independence for several events $A_1, A_2, \ldots, A_n, \ldots \in \mathcal{A}$ is the following: For any subset $(i_1, i_2, ..., i_k)$ of the set of indices $(1, 2, \ldots, n)$ we must have

$$\mu(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = \mu(A_{i_1})\mu(A_{i_2}) \ldots \mu(A_{i_k}) \tag{9}$$

  which means that the joint occurrence of any subset of the events A_i factorizes.

- Important: it is not sufficient to check statistical independence by just considering all possible pairs A_i, A_j of events

- An immediate consequence of definition (7) of statistical independence is Bayes theorem

$$\mu(A_1|A_2) = \mu(A_2|A_1)\frac{\mu(A_1)}{\mu(A_2)} \tag{10}$$

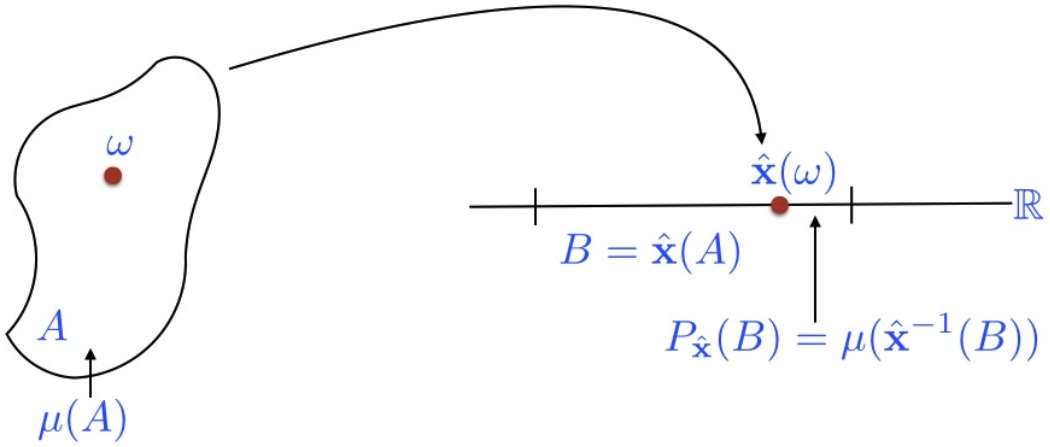

Figure 1. Illustration of the definition of a random variable. A random variable $\hat{\mathbf{x}}$ is a map from the sample space to the set of real numbers. The probability that the random number falls into some Borel set $\mathcal{B}$ is equal to the probability measure $\mu(A)$ of the event $A = \hat{\mathbf{x}}^{-1}(B)$ given by the pre-image of B.

2. Random variables

- The elements ω of the sample space Ω can be rather abstract objects. In practice one often wishes to deal with simple numbers (integer, real or complex numbers) instead of these abstract objects.
- For example, one would like to add and multiply these numbers, and also to consider arbitrary functions of them. The aim is thus to associate numbers with the elements of the sample space. This idea leads to the concept of a *random variable*.

2.1. Definition of random variables

- A random variable $\hat{\mathbf{x}}$ is defined to be a map

$$\hat{\mathbf{x}} : \Omega \to \mathbb{R} \tag{11}$$

  which assigns to each elementary event $\omega \in \Omega$ a real number $\hat{\mathbf{x}}(\omega)$

- Given some $\omega \in \Omega$, the value $x = \hat{\mathbf{x}}(\omega)$ is called a realization of $\hat{\mathbf{x}}$

- To completely define a random variable, we still need to impose a certain condition on the function $\hat{\mathbf{x}}$. To formulate this condition we introduce the σ-algebra of Borel sets of $\mathbb{R}$ which will be denoted by $\mathcal{B}$.

- The σ-algebra of Borel sets of $\mathbb{R}$ ($\mathcal{B}$) is the smallest σ-algebra which contains all subsets of the form $(-\infty, x)$, $x \in \mathbb{R}$. In particular, it contains all open and closed intervals of the real axis.

- Condition: the function $\hat{\mathbf{x}}$ must be a measurable function. This means that for any Borel set $B \in \mathcal{B}$ the pre-image $A = \hat{\mathbf{x}}^{-1}(B)$ belongs to the σ-algebra $\mathcal{A}$ of events.

- This condition ensures that the probability of $\hat{\mathbf{x}}^{-1}(B)$ is well defined and that we can define the probability distribution of $\hat{\mathbf{x}}$ by means of the formula

$$P_{\hat{\mathbf{x}}}(B) = \mu(\hat{\mathbf{x}}^{-1}(B)) \tag{12}$$

- A random variable $\hat{\mathbf{x}}$ thus gives rise to a probability distribution $P_{\hat{\mathbf{x}}}(B)$ on the Borel sets B of the real axis (see Fig. 1)

- Particular Borel sets are the sets $(-\infty, x]$ with $x \in \mathbb{R}$. Consider the pre-images of these set, that is the sets

$$A_x = \{\omega \in \Omega | \hat{\mathbf{x}}(\omega) \leq x\} . \tag{13}$$

By the condition on the map $\hat{\mathbf{x}}$ these sets are measurable for any $x \in \mathbb{R}$. This enables us to introduce the function

$$F_{\hat{\mathbf{x}}}(x) \equiv \mu(A_x) = \mu(\{\omega \in \Omega | \hat{\mathbf{x}}(\omega) \leq x\}) \,. \tag{14}$$

For a given x this function yields the probability that the random number $\hat{\mathbf{x}}$ takes on a value in the interval $(-\infty, x]$. The function $F_{\hat{\mathbf{x}}}(x)$ is referred to as the cumulative distribution function (cdf) of $\hat{\mathbf{x}}$.

- The random variable $\hat{\mathbf{x}}$ is said to have a probability density function (pdf) $f_{\hat{\mathbf{x}}}(x)$ if the cumulative distribution function $F_{\hat{\mathbf{x}}}(x)$ can be represented as

$$F_{\hat{\mathbf{x}}}(x) = \int_{-\infty}^{x} f_{\hat{\mathbf{x}}}(x)\, dx \,. \tag{15}$$

Moreover, if $F_{\hat{\mathbf{x}}}(x)$ is absolutely continuous[1] we get the formula

$$f_{\hat{\mathbf{x}}}(x) = \frac{dF_{\hat{\mathbf{x}}}(x)}{dx} \tag{16}$$

† Definition: Let I be an interval in the real line $\mathbb{R}$. A function $f \colon I \to \mathbb{R}$ is absolutely continuous on I if for every positive number ε, there is a positive number δ such that whenever a finite sequence of pairwise disjoint sub-intervals (x_k, y_k) of I with $x_k, y_k \in I$ satisfies $\sum_k (y_k - x_k) < \delta$, then $\sum_k |f(y_k) - f(x_k)| < \varepsilon$.

A continuous function fails to be absolutely continuous if it fails to be uniformly continuous, which can happen if the domain of the function is not compact – examples are $\tan(x)$ over $[0, \pi/2)$, x^2 over the entire real line, and $\sin(1/x)$ over $(0, 1]$. But a continuous function f can fail to be absolutely continuous even on a compact interval. It may not be "differentiable almost everywhere" (like the Weierstrass function, which is not differentiable anywhere). Or it may be differentiable almost everywhere and its derivative f' may be Lebesgue integrable, but the integral of f' differs from the increment of f (how much f changes over an interval). This happens for example with the Cantor function.

- The pdf $f_{\hat{\mathbf{x}}}(x)$ is one of the most important concepts in the theory of random variables. To be able to consider $f_{\hat{\mathbf{x}}}(x)$ as a *bona fide* pdf, it must satisfy the nonnegativity and normalization conditions

$$f_{\hat{\mathbf{x}}}(x) \geq 0 \quad ; \int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}}(x)\, dx = 1 \tag{17}$$

- The probability that the random variable $\hat{\mathbf{x}}$ takes a value in a finite interval $[a, b]$ is

$$P(\hat{\mathbf{x}} \in [a, b]) = \int_a^b f_{\hat{\mathbf{x}}}(x)\, dx\,. \tag{18}$$

  Indeed, the interpretation of the pdf is that, in the limit $dx \to 0$, $f_{\hat{\mathbf{x}}}(x)dx$ gives the probability that the random variable $\hat{\mathbf{x}}$ takes values between x and $x + dx$, i.e.

$$P(x \leq \hat{\mathbf{x}} \leq x + dx) = f_{\hat{\mathbf{x}}}(x)\, dx \tag{19}$$

  In general, the probability that the probability that a random variable $\hat{\mathbf{x}}$ takes a value within an arbitrary region $\Gamma \subset \mathbb{R}$ of the real numbers is

$$P(\hat{\mathbf{x}} \in \Gamma) = \int_\Gamma f_{\hat{\mathbf{x}}}(x)\, dx \tag{20}$$

- Dimensions: Note that $f_{\hat{\mathbf{x}}}(x)$ has units of the inverse of the units of x, and it is not limited to taking values smaller than or equal to 1. For instance, the pdf governing the probability of the next emission of an electron by a β-radioactive substance has units of inverse of time, or t^{-1}.

- Pdf from frequencies: A pdf can be computed from the experimental data. We first generate M data of the random variable $\hat{\mathbf{x}}$ repeating the experiment M times and recording the outcomes $\{x_1, x_2, \ldots, x_M\}$.

We choose an interval Δx and count the number of times $n(x, x+\Delta x)$ in which the random variable has taken values in the interval $(x, x+\Delta x)$. According to the interpretation of $f_{\hat{\mathbf{x}}}(x)$, this pdf can be estimated as

$$f_{\hat{\mathbf{x}}}(x) \approx \frac{n(x, x+\Delta x)}{\Delta x} \tag{21}$$

A good estimate for $f_{\hat{\mathbf{x}}}(x)$ requires M to be large and Δx to be small. An important issue in probability theory is to be able to conclude whether the observed frequencies are indeed compatible, within unavoidable statistical errors, with the postulated probabilities.

- According to Eq. (15), the cumulative distribution function (or cdf) $F_{\hat{\mathbf{x}}}(x)$ is nothing but the probability that the random variable $\hat{\mathbf{x}}$ takes values less or equal than x, i.e.

$$P(\hat{\mathbf{x}} \leq x) = F_{\hat{\mathbf{x}}}(x) \tag{22}$$

Note also that

$$P(x_1 < \hat{\mathbf{x}} \leq x_2) = F_{\hat{\mathbf{x}}}(x_2) - F_{\hat{\mathbf{x}}}(x_1) \tag{23}$$

- General properties of the cdf $F_{\hat{\mathbf{x}}}(x)$ which derive from the non-negativity and normalization conditions for the pdf $f_{\hat{\mathbf{x}}}(x)$ are:

$$F_{\hat{\mathbf{x}}}(x) \geq 0 \tag{24}$$

$$\lim_{x\to-\infty} F_{\hat{\mathbf{x}}}(x) = 0 \tag{25}$$

$$\lim_{x\to+\infty} F_{\hat{\mathbf{x}}}(x) = 1 \tag{26}$$

$$x_2 > x_1 \Rightarrow F_{\hat{\mathbf{x}}}(x_2) \geq F_{\hat{\mathbf{x}}}(x_1) \tag{27}$$

The last property tells us that $F_{\hat{\mathbf{x}}}(x)$ is a nondecreasing function of its argument.

- It is possible to treat discrete variables in the language of pdfs if we use the Dirac delta function $\delta(x)$ [2]. When the random variable takes a discrete (maybe infinite numerable) set of values $\hat{\mathbf{x}} \in \{x_1, x_2, x_3, \ldots\}$ such that the value x_i has probability p_i, then the pdf can be considered as a sum of Dirac delta functions

$$f_{\hat{\mathbf{x}}}(x) = \sum_{\forall i} p_i \delta(x - x_i) \tag{28}$$

because now $P(\hat{\mathbf{x}} = x_i) = \lim_{\Delta x \to 0} \int_{x_i - \Delta x}^{x_i + \Delta x} f_{\hat{\mathbf{x}}}(x)\, dx = p_i$. The corresponding cumulative distribution function (cdf) is a sum of Heaviside step functions

$$F_{\hat{\mathbf{x}}}(x) = \sum_{\forall i} p_i \theta(x - x_i) \tag{29}$$

with the usual definition

$$\theta(x) = \begin{cases} 0 & x < 0\,, \\ 1 & x \geq 0\,. \end{cases} \tag{30}$$

‡ This mathematical object is not a proper function, but a *distribution* or *generalized function*. This is thought of not as a function itself, but only in relation to how it affects other functions when it is *integrated against them*. In keeping with this philosophy, to define the delta function properly, it is enough to say what the *integral* of the delta function against a sufficiently *good* test function is. In this sense the δ-function is a generalized function or distribution on the real number line that is zero everywhere except at zero, with an integral of one over the entire real line. It can be understood as the limit of a succession of functions $\delta_n(x)$ such that $\delta_n(x)$ decays to zero outside a region of width $1/n$ around $x = 0$ such that the integral $\int_{-\infty}^{\infty} dx\, \delta_n(x) = 1$. One example is $\delta_n(x) = n\mathrm{e}^{-n^2x^2/2}/\sqrt{2\pi}$.

3. Average values, moments and characteristic function

- First note that arbitrary functions of a random variable are also random variables themselves.
- Indeed, as a random variable $\hat{\mathbf{x}}$ assigns a real number $\hat{\mathbf{x}}(\xi)$ to the result of the experiment ξ, it is possible to use a given real function $H(x)$ to define a new random variable $\hat{\mathbf{H}}$ as $\hat{\mathbf{H}}(\xi) = H(\hat{\mathbf{x}}(\xi))$. One defines the average or expected value $E[\hat{\mathbf{H}}]$ of this random variable as

$$E[\hat{\mathbf{H}}] = \int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}}(x) H(x)\, dx \tag{31}$$

  Alternative, very common notations for the average value are $\langle \hat{\mathbf{H}} \rangle$ or simply $E[H]$ and $\langle H \rangle$.

- For a discrete random variable with pdf given by Eq. (28), the average value is

$$E[\hat{\mathbf{H}}] = \sum_{\forall i} p_i H(x_i) \tag{32}$$

- Some important expected values are
  - Mean or average value of the random variable: $E[\hat{\mathbf{x}}] = \langle \hat{\mathbf{x}} \rangle$
  - Moments of order n: $E[\hat{\mathbf{x}}^n] = \langle \hat{\mathbf{x}}^n \rangle$
  - Central moments of order n: $E[(\hat{\mathbf{x}} - \langle \hat{\mathbf{x}} \rangle)^n] = \langle (\hat{\mathbf{x}} - \langle \hat{\mathbf{x}} \rangle)^n \rangle$
  - Variance: $\sigma^2[\hat{\mathbf{x}}] = E[(\hat{\mathbf{x}} - \langle \hat{\mathbf{x}} \rangle)^2] = \langle (\hat{\mathbf{x}} - \langle \hat{\mathbf{x}} \rangle)^2 \rangle$
  - Standard deviation: $\sigma[\hat{\mathbf{x}}]$

- The significance of the variance stems from its property to be a measure for the fluctuations of the random variable $\hat{\mathbf{x}}$, that is, for the extent of deviations of the realizations of $\hat{\mathbf{x}}$ from the mean value $\langle\hat{\mathbf{x}}\rangle$. This fact is expressed, for example, by the Chebyshev inequality which states that the variance controls the probability for such deviations, namely for all $\epsilon > 0$ we have

$$\text{Prob}(|\hat{\mathbf{x}} - \langle\hat{\mathbf{x}}\rangle| > \epsilon) < \frac{1}{\epsilon^2}\sigma^2[\hat{\mathbf{x}}] \tag{33}$$

  In particular, if the variance vanishes then the random number $\hat{\mathbf{x}}$ is, in fact, deterministic, i.e. it takes on the single value $x = \langle\hat{\mathbf{x}}\rangle$ with probability 1. The variance plays an important role in the statistical analysis of experimental data, where it is used, for example, to estimate the standard error of the mean for a sample of realizations obtained in an experiment.

- For a multivariate random variable $\underline{\hat{\mathbf{x}}} = (\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_d)$ one defines the matrix elements of the covariance matrix by

$$\text{cov}[\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j] = E[(\hat{\mathbf{x}}_i - \langle\hat{\mathbf{x}}_i\rangle)\,(\hat{\mathbf{x}}_j - \langle\hat{\mathbf{x}}_j\rangle)] \tag{34}$$

  The $d \times d$ matrix with these coefficients is symmetric and positive semidefinite.

- Transformation of random variables: If two random variables $\hat{\mathbf{y}}$ and $\hat{\mathbf{x}}$ are related by a known function $\hat{\mathbf{y}} = g(\hat{\mathbf{x}})$, then their respective pdfs are also related

$$f_{\hat{\mathbf{y}}}(y) = \sum_m \frac{f_{\hat{\mathbf{x}}}(x_m)}{|\frac{dg}{dx}|_{x=x_m}} \tag{35}$$

  where x_m are the solutions of the equation $y = g(x)$. Indeed, since $\hat{\mathbf{y}}$ and $\hat{\mathbf{x}}$ are directly related by the function $g(x)$, conservation of probability under changes of variables implies that $f_{\hat{\mathbf{y}}}(y)|dy| = \sum_m f_{\hat{\mathbf{x}}}(x_m)|dx|_{x_m}$, and from this the above expression immediately follows.

- Example: if the change is $\hat{\mathbf{y}} = \hat{\mathbf{x}}^2$ then the equation $y = x^2$ has no solutions for $y < 0$ and two solutions $x = +y$, $x = -y$ for $y \geq 0$,and the pdf for $\hat{\mathbf{y}}$ is

$$f_{\hat{\mathbf{y}}}(y) = \begin{cases} 0 & y < 0\,, \\ \dfrac{f_{\hat{\mathbf{x}}}(\sqrt{y}) + f_{\hat{\mathbf{x}}}(-\sqrt{y})}{2\sqrt{y}} & y \geq 0\,. \end{cases} \tag{36}$$

- Characteristic function: Let us finally introduce a further important expectation value which may serve to characterize completely a random variable. This is the *characteristic function* or *moment generating function* which is defined as the Fourier transform of the probability density

$$G(k) = \langle \mathrm{e}^{ik\hat{\mathbf{x}}} \rangle = \int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}}(x)\, \mathrm{e}^{ikx}\, dx \tag{37}$$

  It can be shown that the characteristic function G(k) uniquely determines the corresponding probability distribution of $\hat{\mathbf{x}}$.

- Moments of the pdf from the characteristic function: Under the condition that the moments of $\hat{\mathbf{x}}$ exist, the derivatives of $G(k)$ evaluated at $k = 0$ yield the moments of $\hat{\mathbf{x}}$

$$E[\hat{\mathbf{x}}^n] = \langle \hat{\mathbf{x}}^n \rangle = \frac{1}{\mathrm{i}^n} \frac{d^n G(k)}{dk^n}\Big|_{k=0} \tag{38}$$

  This expression immediately follows from the definition of the characteristic function above. This is the reason why the characteristic function is also known as the moment generating function.

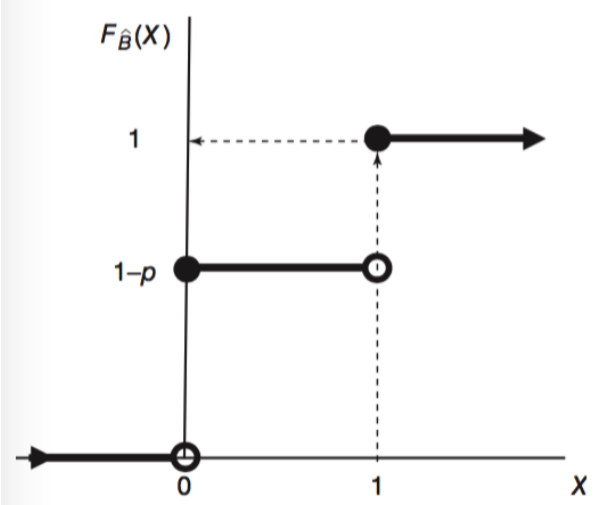

Figure 2. Cumulative distribution function (cdf) $F_{\hat{\mathbf{B}}}(x)$ of the Bernoulli random variable $\hat{\mathbf{B}}(p)$.

4. Some Important Probability Distributions

- In this section we review some of the most usual probability distribution functions found in physics and other branches of science, and summarize their main properties.

4.1. Bernuilli distribution

- It describes a binary experiment in which only two exclusive options are possible: A or $\bar{A}$ ("heads or tails", "either it rains or not"), with respective probabilities p and $1-p$, being $p \in [0,1]$.

- We define the discrete Bernuilli random variable $\hat{\mathbf{B}}(\xi)$ as taking the value 1 (respectively 0) if the

experiment yields $\xi = A$ (respectively $\xi = \bar{A}$). The probabilities are

$$f_{\hat{\mathbf{B}}}(k) = \begin{cases} p & k = 1\,, \\ 1-p & k = 0\,. \end{cases} \tag{39}$$

- We can write the distribution $f_{\hat{\mathbf{B}}}$ as continuous pdf using Dirac-delta functions

$$f_{\hat{\mathbf{B}}}(x) = p\delta(x-1) + (1-p)\delta(x) \tag{40}$$

- Mean value and variance

$$E[\hat{\mathbf{B}}] = \sum_{k=0,1} k\, f_{\hat{\mathbf{B}}}(k) = p\,, \tag{41}$$

$$\sigma^2[\hat{\mathbf{B}}] = \sum_{k=0,1} (k-p)^2\, f_{\hat{\mathbf{B}}}(k) = p(1-p)\,. \tag{42}$$

- When needed below, we will use the notation $\hat{\mathbf{B}}(p)$ to denote a random variable that follows a Bernoulli distribution with parameter p

- The Bernuilli cumulative distribution function now reads

$$F_{\hat{\mathbf{B}}}(x) = \int_{-\infty}^{x} f_{\hat{\mathbf{B}}}(y)\, dy = \begin{cases} 0 & x < 0\,, \\ 1-p & 0 \le x < 1\,, \\ 1 & x \ge 1\,. \end{cases} \tag{43}$$

This is plotted in Fig. 2.

4.2. Binomial distribution

- We now repeat M times the binary experiment of the previous case and count how many times A appears (independently of the order of appearance). This defines a random variable which we call $\hat{\mathbf{N}}_B$. It is a discrete variable that can take any integer value between 0 and M with probabilities

$$f_{\hat{\mathbf{N}}_B}(n) = \binom{M}{n} p^n (1-p)^{M-n} \tag{44}$$

- The random variable $\hat{\mathbf{N}}_B$ is said to follow a binomial distribution. The mean value and variance are given by

$$E[\hat{\mathbf{N}}_B] = Mp\,, \tag{45}$$
$$\sigma^2[\hat{\mathbf{N}}_B] = Mp(1-p)\,. \tag{46}$$

- For the proof, we use the following properties:

$$n\binom{M}{n} = n\frac{M!}{n!(M-n)!} = M\frac{(M-1)!}{(n-1)![(M-1)-(n-1)]!} = M\binom{M-1}{n-1} \tag{47}$$

$$\sum_{n=0}^{M} \binom{M}{n} p^n (1-p)^{M-n} = 1\,. \tag{48}$$

Hence, using these properties

$$
\begin{aligned}
E[\hat{\mathbf{N}}_B] &= \sum_{n=0}^{M} n \binom{M}{n} p^n (1-p)^{M-n} = Mp \sum_{n=1}^{M} \binom{M-1}{n-1} p^{n-1} (1-p)^{(M-1)-(n-1)} \\
&= Mp \sum_{n'=0}^{M'} \binom{M'}{n'} p^{n'} (1-p)^{M'-n'} = Mp\,, \\
E[\hat{\mathbf{N}}_B^2] &= \sum_{n=0}^{M} n^2 \binom{M}{n} p^n (1-p)^{M-n} = Mp \sum_{n=1}^{M} n \binom{M-1}{n-1} p^{n-1} (1-p)^{(M-1)-(n-1)} \\
&= Mp \sum_{n'=0}^{M'} (n'+1) \binom{M'}{n'} p^{n'} (1-p)^{M'-n'} = Mp\,[(M-1)p+1] \qquad (49) \\
\sigma^2[\hat{\mathbf{N}}_B] &= E[\hat{\mathbf{N}}_B^2] - E[\hat{\mathbf{N}}_B]^2 = Mp + M(M-1)p^2 - M^2p^2 = Mp(1-p)
\end{aligned}
$$

- We can also derive the mean and variance from the characteristic function of the binomial pdf

$$
G_{\hat{\mathbf{N}}_B}(k) = \sum_{n=0}^{M} \mathrm{e}^{\mathrm{i}kn} f_{\hat{\mathbf{N}}_B}(n) = \sum_{n=0}^{M} \binom{M}{n} (\mathrm{e}^{\mathrm{i}k} p)^n (1-p)^{M-n} = \left[\mathrm{e}^{\mathrm{i}k} p + 1 - p\right]^M \qquad (50)
$$

Using now Eq. (38) to write the moments of the binomial pdf in terms of the derivatives of its

characteristic function, we simply find

$$
\begin{aligned}
E[\hat{\mathbf{N}}_B] &= \frac{1}{\mathrm{i}}\frac{dG(k)}{dk}|_{k=0} = \frac{M}{\mathrm{i}}\left[(\mathrm{e}^{\mathrm{i}k}p+1-p)^{M-1}\mathrm{i}p\mathrm{e}^{\mathrm{i}k}\right]_{k=0} = Mp \\
E[\hat{\mathbf{N}}_B^2] &= \frac{1}{\mathrm{i}^2}\frac{d^2G(k)}{dk^2}|_{k=0} \\
&= \frac{Mp}{\mathrm{i}}\left[\mathrm{i}\mathrm{e}^{\mathrm{i}k}(\mathrm{e}^{\mathrm{i}k}p+1-p)^{M-1} + (M-1)(\mathrm{e}^{\mathrm{i}k}p+1-p)^{M-2}\mathrm{i}p\mathrm{e}^{\mathrm{i}2k}\right]_{k=0} \\
&= Mp\left[1+(M-1)p\right] \\
\sigma^2[\hat{\mathbf{N}}_B] &= E[\hat{\mathbf{N}}_B^2] - E[\hat{\mathbf{N}}_B]^2 = Mp + M(M-1)p^2 - M^2p^2 = Mp(1-p)
\end{aligned}
$$

- We will denote by $\hat{\mathbf{N}}_B(p, M)$ a random variable that follows a binomial distribution with probability p and number of repetitions M.

4.3. Geometric distribution

- We consider, again, repetitions of the binary experiment, but now the random variable $\hat{\mathbf{N}}_G$ is defined as the number of times we must repeat the experiment before the result A appears
- This is a discrete random variable that can take any integer value $0, 1, 2, 3, \ldots$ The probability that it takes a value equal to n is

$$
f_{\hat{\mathbf{N}}_G}(n) = (1-p)^n p\,, \quad n = 0, 1, 2, \ldots \tag{51}
$$

- Mean value and variance

$$E[\hat{\mathbf{N}}_G] = \frac{1-p}{p}\,, \tag{52}$$

$$\sigma^2[\hat{\mathbf{N}}_G] = \frac{1-p}{p^2}\,. \tag{53}$$

- For the proof, we just need the geometric series result (which yields the normalization of the above pdf)

$$\sum_{n=0}^{\infty} q^n = \frac{1}{1-q}\,, \qquad \text{for } |q| \leq 1 \tag{54}$$

Using this it is simple to show that

$$\begin{aligned}
\langle \hat{\mathbf{N}}_G \rangle &= \sum_{n=1}^{\infty} n\,(1-p)^n p = (1-p)\sum_{n'=0}^{\infty}(n'+1)\,(1-p)^{n'} p \\
&= (1-p)(\langle \hat{\mathbf{N}}_G \rangle + 1) \qquad \Rightarrow \qquad \langle \hat{\mathbf{N}}_G \rangle = \frac{1-p}{p}\,, \\
\langle \hat{\mathbf{N}}_G^2 \rangle &= \sum_{n=1}^{\infty} n^2\,(1-p)^n p = (1-p)\sum_{n'=0}^{\infty}(n'+1)^2\,(1-p)^{n'} p \\
&= (1-p)(\langle \hat{\mathbf{N}}_G^2 \rangle + 2\langle \hat{\mathbf{N}}_G \rangle + 1) \qquad \Rightarrow \qquad \langle \hat{\mathbf{N}}_G^2 \rangle = \frac{2-3p+p^2}{p^2}\,, \\
\sigma^2[\hat{\mathbf{N}}_G] &= \langle \hat{\mathbf{N}}_G^2 \rangle - \langle \hat{\mathbf{N}}_G \rangle^2 = \frac{2-3p+p^2-(1-p)^2}{p^2} = \frac{1-p}{p^2}\,.
\end{aligned}$$

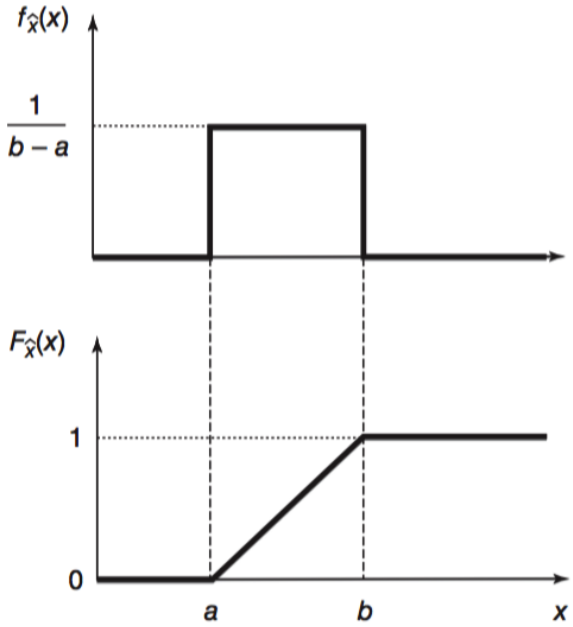

Figure 3. Cumulative distribution function (cdf) $F_{\hat{\mathbf{U}}}(x)$ of the uniform random variable $\hat{\mathbf{U}}(a,b)$.

- The characteristic function for the geometric pdf reads

$$G_{\hat{\mathbf{N}}_G}(k) = \sum_{n=0}^{M} \mathrm{e}^{\mathrm{i}kn} f_{\hat{\mathbf{N}}_G}(n) = \sum_{n=0}^{M} [\mathrm{e}^{\mathrm{i}k}(1-p)]^n p = \frac{p}{1-\mathrm{e}^{\mathrm{i}k}(1-p)} \,. \tag{55}$$

  The above moments can be simply obtained by deriving this expression with respect to k.

4.4. Uniform distribution

- This is our first example of a continuous random variable. We want to describe an experiment in which all possible results are real numbers within the interval (a, b) occurring with the same probability, while no result can appear outside this interval.

- We will use the notation $\hat{\mathbf{U}}(a, b)$ to indicate a uniform random variable in the interval (a, b). The pdf is then constant within the interval (a, b) and 0 outside it. Applying the normalization condition, it is precisely

$$f_{\hat{\mathbf{U}}}(x) = \begin{cases} \dfrac{1}{b-a} & x \in [a, b]\,, \\ 0 & x \notin [a, b]\,. \end{cases} \tag{56}$$

- The cumulative distribution function is

$$F_{\hat{\mathbf{U}}}(x) = \int_{-\infty}^{x} f_{\hat{\mathbf{U}}}(y)\, dy = \begin{cases} 0 & x < a\,, \\ \dfrac{x-a}{b-a} & a \le x < b\,, \\ 1 & x > b\,. \end{cases} \tag{57}$$

  This two functions are plotted in Fig. 3.

- The average and the variance now read

$$E[\hat{\mathbf{U}}] = \int_{-\infty}^{\infty} x\, f_{\hat{\mathbf{U}}}(x)\, dx = \frac{a+b}{2}\,, \tag{58}$$

$$\sigma^2[\hat{\mathbf{U}}] = \int_{-\infty}^{\infty} x^2\, f_{\hat{\mathbf{U}}}(x)\, dx - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}\,. \tag{59}$$

- Characteristic function for the uniform pdf

$$G_{\hat{\mathbf{U}}}(k) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}kx}\, f_{\hat{\mathbf{U}}}(x)\, dx = \frac{1}{b-a}\int_a^b \mathrm{e}^{\mathrm{i}kx}\, dx = \frac{\mathrm{i}(\mathrm{e}^{\mathrm{i}ka} - \mathrm{e}^{\mathrm{i}kb})}{k(b-a)} \tag{60}$$

- The uniform distribution $\hat{\mathbf{U}}(0,1)$ appears in an important result. Let us consider an arbitrary random variable $\hat{\mathbf{x}}$ (discrete or continuous) whose cdf is $F_{\hat{\mathbf{x}}}(x)$, that we assume invertible, and let us define the new random variable $\hat{\mathbf{u}} = F_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})$. We will prove now that $\hat{\mathbf{u}} = F_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})$ is a $\hat{\mathbf{U}}(0,1)$ variable.

- To prove this result, we study the cdf $F_{\hat{\mathbf{u}}}(u)$ of the random variable $\hat{\mathbf{u}}$,

$$F_{\hat{\mathbf{u}}}(u) = P(\hat{\mathbf{u}} \le u) = P(F_{\hat{\mathbf{x}}}(x) \le u)\,. \tag{61}$$

  First note that $F_{\hat{\mathbf{x}}}(x) \in [0,1]$ as for any cumulative distribution function. This immediately implies that for $u < 0$ we have $P(F_{\hat{\mathbf{x}}}(x) \le u) = 0$, while for $u > 1$ we have $P(F_{\hat{\mathbf{x}}}(x) \le u) = 1$. Now, for $u \in [0,1]$, the condition $F_{\hat{\mathbf{x}}}(x) \le u$ is equivalent to $\hat{\mathbf{x}} \le F^{-1}_{\hat{\mathbf{x}}}(u)$ (recall we assume $F_{\hat{\mathbf{x}}}(x)$ to be invertible). Therefore

$$F_{\hat{\mathbf{u}}}(u) = P(F_{\hat{\mathbf{x}}}(x) \le u) = P(\hat{\mathbf{x}} \le F^{-1}_{\hat{\mathbf{x}}}(u)) = F_{\hat{\mathbf{x}}}(F^{-1}_{\hat{\mathbf{x}}}(u)) = u. \tag{62}$$

  Therefore we have found that

$$F_{\hat{\mathbf{u}}}(u) = \begin{cases} 0 & u < 0\,, \\ u & 0 \le u < 1\,, \\ 1 & u > 1\,, \end{cases} \tag{63}$$

  which is nothing but the cdf of a uniform random variable $\hat{\mathbf{U}}(0,1)$.

4.5. Poisson distribution

- Let us consider the binomial distribution of Subsection 4.2 above in the limit of infinitely many repetitions M. If we take the double limit $M \to \infty$, $p \to 0$ but keeping $Mp \to \lambda$, with λ a finite value, the binomial distribution $\hat{\mathbf{N}}_B(p, M)$ tends to the so-called Poisson distribution $\hat{\mathbf{P}}(\lambda)$.

- The form of the Poisson distribution $f_{\hat{\mathbf{P}}}(n)$ now follows from Stirling formula[1] $m! \approx m^m \mathrm{e}^{-m}\sqrt{2\pi m}$ applied to the binomial distribution $f_{\hat{\mathbf{N}}_B}(n)$ in Eq. (44). For the binomial coefficient

$$\binom{M}{n} = \frac{M!}{n!(M-n)!} \approx \frac{M^M \mathrm{e}^{-M}\sqrt{2\pi M}}{n!(M-n)^{M-n}\mathrm{e}^{-(M-n)}\sqrt{2\pi(M-n)}} = \sqrt{\frac{M}{M-n}}\frac{M^M \mathrm{e}^{-n}}{n!(M-n)^{M-n}}$$

† Stirling formula is easily derived from the Gamma function representation of the factorial. Indeed,

$$n! = \Gamma(n+1) = \int_0^\infty x^n \mathrm{e}^{-x}\, dx = \int_0^\infty e^{n\ln x - x}\, \mathrm{d}x = n e^{n\ln n}\int_0^\infty e^{n(\ln y - y)}\, \mathrm{d}y.$$

where we have applied the change of variables $y = x/n$ in the last equality. The last integral can be approximated using Laplace (or steepest descent) method

$$\int_0^\infty e^{nf(y)}\, dy \approx e^{nf(y_0)}\int_0^\infty e^{-n|f''(y_0)|(y-y_0)^2/2}\, dy \approx \sqrt{\frac{2\pi}{n|f''(y_0)|}} e^{nf(y_0)} \text{ as } n \to \infty. \tag{64}$$

with y_0 solution of $f'(y_0) = 0$. For $f(y) = \ln y - y$ (note that then $y_0 = 1$), this yields the desired Stirling approximation

$$n! \approx n^n \mathrm{e}^{-n}\sqrt{2\pi n}\,.$$

Using this in the binomial pdf, neglecting subleading terms, and substituting $p \to \lambda/M$,

$$f_{\hat{\mathbf{N}}_B}(n) = \binom{M}{n} p^n (1-p)^{M-n} \approx \frac{M^M p^n (1-p)^{M-n} \mathrm{e}^{-n}}{n!(M-n)^{M-n}} = \frac{M^M (\frac{\lambda}{M})^n \left(1-\frac{\lambda}{M}\right)^{M-n} \mathrm{e}^{-n}}{n! M^{M-n} \left(1-\frac{n}{M}\right)^{M-n}}$$

$$= \frac{\lambda^n \left(1-\frac{\lambda}{M}\right)^{M-n} \mathrm{e}^{-n}}{n! \left(1-\frac{n}{M}\right)^{M-n}} \approx \frac{\lambda^n \left(1-\frac{\lambda}{M}\right)^{M} \mathrm{e}^{-n}}{n! \left(1-\frac{n}{M}\right)^{M}} .$$

Now, as $M \to \infty$, we have that $(1-\frac{x}{M})^M \to \mathrm{e}^{-x}$, so finally

$$\binom{M}{n} p^n (1-p)^{M-n} \approx \frac{\lambda^n}{n!} \mathrm{e}^{-\lambda} \equiv f_{\hat{\mathbf{P}}}(n) , \qquad n = 0, 1, 2, \ldots, \infty . \tag{65}$$

- The Poisson distribution is one of the most important distributions in nature, probably second only to the Gaussian distribution (to be discussed later).

- Mean value and variance

$$E[\hat{\mathbf{P}}] = \sigma^2[\hat{\mathbf{P}}] = \lambda \tag{66}$$

The equality of the mean and variance is a typical property characterizing the Poisson distribution.

- Proof:

$$E[\hat{\mathbf{P}}] = \sum_{n=1}^{\infty} n \frac{\lambda^n}{n!} \mathrm{e}^{-\lambda} = \sum_{n'=0}^{\infty} (n'+1) \frac{\lambda}{n'+1} \frac{\lambda^{n'}}{n'!} \mathrm{e}^{-\lambda} = \lambda \sum_{n'=0}^{\infty} \frac{\lambda^{n'}}{n'!} \mathrm{e}^{-\lambda} = \lambda , \tag{67}$$

$$E[\hat{\mathbf{P}}^2] = \sum_{n=1}^{\infty} n^2 \frac{\lambda^n}{n!} \mathrm{e}^{-\lambda} = \lambda \sum_{n'=0}^{\infty} (n'+1) \frac{\lambda^{n'}}{n'!} \mathrm{e}^{-\lambda} = \lambda (E[\hat{\mathbf{P}}] + 1) = \lambda(\lambda+1) , \tag{68}$$

$$\sigma^2[\hat{\mathbf{P}}] = E[\hat{\mathbf{P}}^2] - E[\hat{\mathbf{P}}]^2 = \lambda(\lambda+1) - \lambda^2 = \lambda . \tag{69}$$

- Characteristic function for the Poisson distribution

$$G_{\hat{\mathbf{P}}}(k) = \sum_{n=0}^{\infty} \frac{(\lambda \mathrm{e}^{\mathrm{i}k})^n}{n!} \mathrm{e}^{-\lambda} = \exp[\lambda(\mathrm{e}^{\mathrm{i}k} - 1)] \tag{70}$$

- Example: A convenient approximation: We can think of the Poisson distribution simply as a convenient limit that simplifies the calculations in many occasions. For instance, the probability that a person was born on a particular day, say 1 January, is $p = 1/365$, approximately. Then the probability of being born any other day is $1 - p = 364/365$. Imagine that we have now a large group of $M = 500$ people. What is the probability that exactly three people were born on 1 January? The correct answer is given by the binomial distribution by considering the events A=$\{$*being born on 1 January*$\}$ with probability $p = 1/365$ and $\bar{A}$=$\{$*not being born on 1 January*$\}$ with probability $1 - p = 364/365$:

$$P(\hat{\mathbf{N}}_B = 3) = \binom{500}{3} \left(\frac{1}{365}\right)^3 \left(\frac{364}{365}\right)^{497} = 0.108919\ldots \tag{71}$$

  As p is small and M large, it seems justified to use the Poisson approximation. In this case, $\lambda = pM \approx 500/365 = 1.37$, so we obtain

$$P(\hat{\mathbf{P}} = 3) = \mathrm{e}^{-1.37} \frac{1.37^3}{3!} = 0.108900\ldots \tag{72}$$

  which yields a reasonably good approximation.

- Example: Exact limit: There are occasions in which the Poisson limit occurs exactly. Imagine we distribute M dots randomly with a distribution $\hat{\mathbf{U}}[0, T]$, uniform in the interval $[0, T]$. We will think immediately of this as events occurring randomly in time with a uniform rate, hence the notation. We

call $\omega = M/T$ the *rate* (or *frequency*) at which points are distributed. We now ask the question: what is the probability that exactly k of the M dots lie in the interval $[t_1, t_1 + t] \in [0, T]$?

- The event A=$\{$*one given dot lies in the interval* $[t_1, t_1 + t]\}$ has probability $p = t/T$, whereas the event $\bar{A}$=$\{$*the given dot does not lie in the interval* $[t_1, t_1 + t]\}$ has probability $q = 1 - p$. The required probability is given by the binomial distribution $\hat{\mathbf{B}}(p, M)$ defined in Eq. (44)

- We now make the limit $M \to \infty$, $T \to \infty$ but $\omega = M/T$ finite. This limit corresponds to the Poisson limit of the binomial pdf explained above, where $M \to \infty$, $p = t/T \to 0$ with $pM = \omega t$ finite. Physically, this limit corresponds to the distribution in which the events occur uniformly in time with a rate (frequency) ω. As mentioned earlier, it can be proven using Stirling's approximation that, in this limit, the binomial distribution $\hat{\mathbf{B}}(p, M)$ tends to a Poisson distribution $\hat{\mathbf{P}}(\lambda)$ of parameter $\lambda = pM = \omega t$, finite.

- Example: β-radiactive decay: Consider N atoms of β-radioactive substance. Each atom emits one electron independently of the others. The probability that the given atom will disintegrate is constant with time, and the number of atoms in the substance is so large (of the order of Avogadro's number) that the relative change of the number of atoms in the substance during the time interval of interest is negligible. We hence can assume a constant decay rate ω which can be estimated simply by counting the number of electrons M emitted in a time interval T as $\omega = M/T$. Under these circumstances, the number k of electrons emitted in the time interval $[t_1, t_1 + t]$ follows a Poisson distribution with parameter $\lambda = pM = tM/T = \omega t$, or

$$P(k; t) = \mathrm{e}^{-\omega t} \frac{(\omega t)^k}{k!} \tag{73}$$

4.6. Exponential distribution

- A continuous random variable $\hat{\mathbf{x}}$ follows an exponential distribution if its pdf is

$$f_{\hat{\mathbf{x}}}(x) = \begin{cases} 0 & x < 0\,, \\ a\mathrm{e}^{-ax} & x \geq 0\,. \end{cases} \tag{74}$$

  with $a > 0$ a parameter.

- The mean value and variance are

$$E[\hat{\mathbf{P}}] = \int_0^\infty x\, a\mathrm{e}^{-ax}\, dx = \frac{\Gamma(2)}{a} = \frac{1}{a}\,, \tag{75}$$

$$E[\hat{\mathbf{P}}^2] = \int_0^\infty x^2\, a\mathrm{e}^{-ax}\, dx = \frac{\Gamma(3)}{a^2} = \frac{2}{a^2}\,, \tag{76}$$

$$\sigma^2[\hat{\mathbf{P}}] = E[\hat{\mathbf{P}}^2] - E[\hat{\mathbf{P}}]^2 = \frac{2}{a^2} - \frac{1}{a^2} = \frac{1}{a^2}\,. \tag{77}$$

- Example: Consider the emission of electrons by a radioactive substance which we know is governed by the Poisson distribution for those time intervals such that the emission rate can be considered constant. Let us set our clock at $t = 0$ and then measure the time t of the *first* observed emission of an electron. This time is a random variable $\hat{\mathbf{t}}$ (a number associated with the result of an experiment) and has a pdf which we call $f^{\rm 1st}_{\hat{\mathbf{t}}}(t)$. By definition, $f^{\rm 1st}_{\hat{\mathbf{t}}}(t)\, dt$ is the probability that the first electron is emitted during the interval $(t, t + dt)$ and, accordingly, the probability that the first electron is emitted after time t is $\int_t^\infty f^{\rm 1st}_{\hat{\mathbf{t}}}(t')\, dt'$. This is equal to the probability that no electrons have been emitted during $(0, t)$, i.e. $P(0; t) = e^{-\omega t}$, see Eq. (73), that is

$$\int_t^\infty f^{\rm 1st}_{\hat{\mathbf{t}}}(t')\, dt' = e^{-\omega t} \quad t \geq 0\,. \tag{78}$$

Taking the time derivate on both sides of this equation, we obtain

$$f_{\hat{\mathbf{t}}}^{1\text{st}}(t) = \omega\, e^{-\omega t} \quad t \geq 0\,. \tag{79}$$

which is nothing but the exponential distribution. The same exponential pdf rules the distribution of time intervals between consecutive events in constant rate problems.

- Alternatively, if $\hat{\mathbf{t}}$ follows the previous exponential distribution, then the number of events occurring in a time interval $(0, 1)$ follows a Poisson $\hat{\mathbf{P}}(\lambda)$ distribution with $\lambda = \omega \times 1 = \omega$.

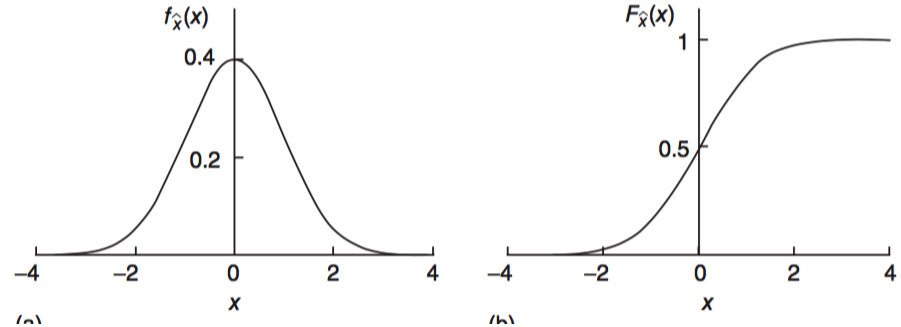

Figure 4. Pdf and cdf of the Gaussian distribution of mean 0 and variance 1.

4.7. Gaussian distribution

- A continuous random variable $\hat{\mathbf{x}}$ follows a Gaussian distribution of mean μ and variance σ^2 if its pdf is

$$f_{\hat{\mathbf{x}}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] . \tag{80}$$

  We will use the notation that $\hat{\mathbf{x}}$ is a $\hat{\mathbf{G}}(\mu, \sigma)$ random variable.

- The average and the variance are, obviously $E[\hat{\mathbf{x}}] = \mu$ and $\sigma^2[\hat{\mathbf{x}}] = \sigma^2$.

- The cdf is

$$F_{\hat{\mathbf{x}}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{(x'-\mu)^2}{2\sigma^2}}\, dx' = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right) \tag{81}$$

  where erf(z) is the error function

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \mathrm{e}^{-y^2}\, dy \tag{82}$$

Fig. 4 displays the Gaussian pdf and cdf.

- Interestingly, the characteristic function of the Gaussian pdf is another Gaussian function

$$G_{\hat{\mathbf{G}}}(k) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}kx} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \frac{\mathrm{e}^{\mathrm{i}k\mu-\sigma^2k^2/2}}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{[x-(\mu+\mathrm{i}k\sigma^2)]^2}{2\sigma^2}}\, dx = \mathrm{e}^{\mathrm{i}k\mu-\sigma^2k^2/2}\,.$$

- Gaussian random variables are very important in practice because they appear in a large number of problems, either as an exact distribution in some limit or, simply, as providing a sufficient approximation to the real distribution. After all, it is not unusual that many distributions have a maximum value and this can in many cases be approximated by a Gaussian distribution (the so-called bell-shaped curve).

- One of the reasons for the widespread appearance of Gaussian distributions is the central-limit theorem, which states that the sum of a large number of independent random variables, whatever their distribution, will approach a Gaussian distribution. See Section 9 below for a detailed discussion.

- One can prove that the binomial distribution $\hat{\mathbf{B}}(p, M)$ tends to the Gaussian distribution $\hat{\mathbf{G}}(Mp, Mp(1-p))$ in the limit $M \to \infty$,

$$f_{\hat{\mathbf{N}}_B}(n) = \binom{M}{n} p^n (1-p)^{M-n} \xrightarrow{M\to\infty} \frac{1}{\sqrt{2\pi Mp(1-p)}} \exp\left[-\frac{(n-Mp)^2}{2Mp(1-p)}\right]. \tag{83}$$

  Note however that the mean and the variance of this Gaussian distribution are related, so only a particular family of Gaussian distribution can be obtained as limits of binomial distributions.

- Normal approximation for the binomial distribution: We now want to prove that $\hat{\mathbf{B}}(p, M) \xrightarrow{M\to\infty} \hat{\mathbf{G}}(Mp, Mp(1-p))$. To prove this result, we will assume that, together with M, also Mp and $M(1-p)$

are large, which is true always for any *fixed* p (i.e. not scaling with M) in the limit of large M. To simplify the formulas below, we now introduce $q = 1 - p$. Using Stirling's formula

$$f_{\hat{\mathbf{N}}_B}(n) \approx \frac{M^M \mathrm{e}^{-M}\sqrt{2\pi M}}{n^n \mathrm{e}^{-n}\sqrt{2\pi n}(M-n)^{M-n}\mathrm{e}^{-(M-n)}\sqrt{2\pi(M-n)}} p^n q^{M-n}$$
$$= \left(\frac{Mp}{n}\right)^n \left(\frac{Mq}{M-n}\right)^{M-n} \sqrt{\frac{M}{2\pi n(M-n)}}\,. \tag{84}$$

It is convenient now to define the excess variable $\delta = n - Mp$ (recall that the binomial distribution hasmean Mp and variance Mpq), so we have $n = \delta + Mp$ and $M - n = Mq - \delta$. Moreover, taking logarithms and recalling that $\ln(1+x) \approx x - \frac{1}{2}x^2 + O(x^3)$,

$$\ln\left(\frac{Mp}{n}\right) = \ln\left(\frac{Mp}{Mp+\delta}\right) = -\ln(1+\frac{\delta}{Mp}) \approx -\frac{\delta}{Mp} + \frac{\delta^2}{2M^2p^2}\,,$$
$$\ln\left(\frac{Mq}{M-n}\right) = \ln\left(\frac{Mq}{Mq-\delta}\right) = -\ln(1-\frac{\delta}{Mq}) \approx \frac{\delta}{Mq} + \frac{\delta^2}{2M^2q^2}\,,$$

In this way

$$\ln\left[\left(\frac{Mp}{n}\right)^n \left(\frac{Mq}{M-n}\right)^{M-n}\right] = (Mp+\delta)\ln\left(\frac{Mp}{Mp+\delta}\right) + (Mq-\delta)\ln\left(\frac{Mq}{Mq-\delta}\right)$$
$$\approx (Mp+\delta)\left(-\frac{\delta}{Mp} + \frac{\delta^2}{2M^2p^2}\right) + (Mq-\delta)\left(\frac{\delta}{Mq} + \frac{\delta^2}{2M^2q^2}\right) \approx -\frac{\delta^2}{2Mpq} + O\left(-\frac{\delta^3}{M^2}\right)\,.$$

Exponentiating the above result, we thus find

$$\left(\frac{Mp}{n}\right)^n \left(\frac{Mq}{M-n}\right)^{M-n} \approx \mathrm{e}^{-\frac{\delta^2}{2Mpq}}\left[1 + O\left(\frac{\delta^3}{M^2}\right)\right] \tag{85}$$

For the square root factor in Eq. (84), we also have

$$\sqrt{\frac{M}{2\pi n(M-n)}} = \sqrt{\frac{M}{2\pi(Mp+\delta)(Mq-\delta)}} \approx \sqrt{\frac{1}{2\pi Mpq}}\left[1+O\left(\frac{\delta}{M}\right)\right] \tag{86}$$

To end this calculation, note that we want to approximate the binomial distribution in the region where it is appreciably distinct from 0, i.e. we want n to be *close* to the average Mp, or in other words no more than a few standard deviations σ from the average. As $\sigma = \sqrt{Mpq} \sim O(\sqrt{M})$, then we will have $\delta = n - Mp \sim O(\sqrt{M})$, and this implies that the corrections

$$O\left(\frac{\delta^3}{M^2}\right) \sim O\left(\frac{\delta}{M}\right) \sim O\left(\frac{1}{\sqrt{M}}\right) \xrightarrow{M\to\infty} 0 \tag{87}$$

Therefore, as stated above,

$$f_{\hat{\mathbf{N}}_B}(n) = \binom{M}{n} p^n (1-p)^{M-n} \xrightarrow{M\to\infty} \frac{1}{\sqrt{2\pi Mp(1-p)}} \exp\left[-\frac{(n-Mp)^2}{2Mp(1-p)}\right] . \tag{88}$$

- As one could easily expect, the Gaussian distribution can also be obtained as the limit of the Poisson distribution for large parameter $\lambda \to \infty$. This yields a *particular* Gaussian distribution of the same mean and variance, or $\hat{\mathbf{G}}(\lambda, \lambda)$. Again, although the exact result refers to the limit $\lambda \to \infty$, in practice the approximation can be considered sufficient for $\lambda \geq 100$, especially around the maximum of the distribution (Figure 5).

- To see this result, we write

$$f_{\hat{\mathbf{P}}}(n) = \frac{\lambda^n}{n!}\mathrm{e}^{-\lambda} \overset{\text{Stirling}}{\approx} \left(\frac{\lambda}{n}\right)^n \frac{\mathrm{e}^{-\lambda+n}}{\sqrt{2\pi n}} = \left(\frac{\lambda}{\lambda+\delta}\right)^{\lambda+\delta} \frac{\mathrm{e}^{\delta}}{\sqrt{2\pi(\lambda+\delta)}}$$

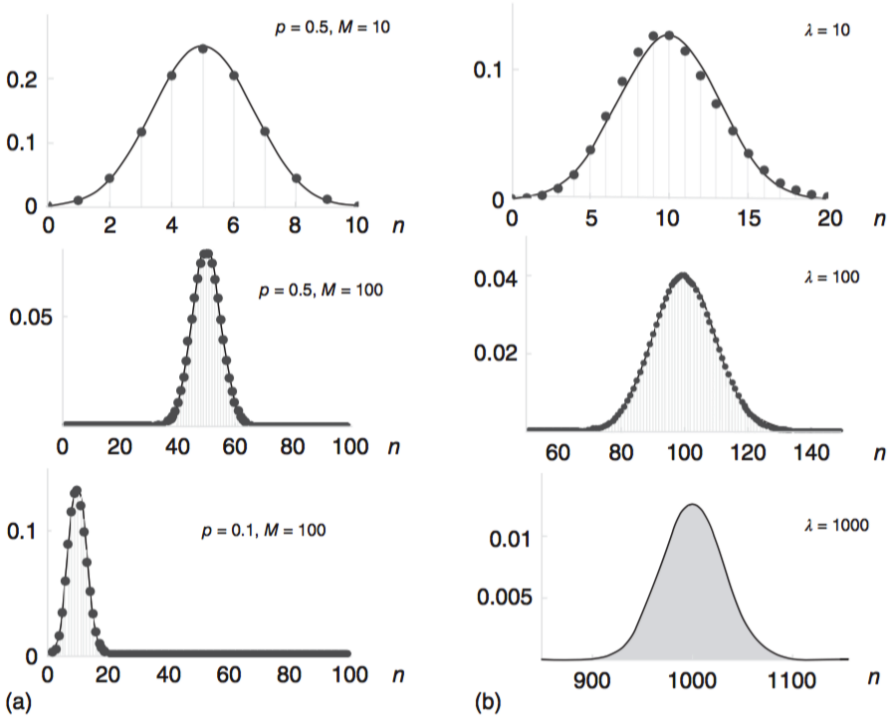

Figure 5. (a) Binomial distribution (dots) and its Gaussian approximation (solid line). (b) Poisson distribution (dots) and its binomial approximation (solid line).

where we have defined as above the excess variable $\delta = n - \lambda$. Using now that

$$\ln\left(\frac{\lambda}{\lambda+\delta}\right)^{\lambda+\delta} = (\lambda+\delta)\ln\left(\frac{\lambda}{\lambda+\delta}\right) \approx -\delta - \frac{\delta^2}{2\lambda}\,, \tag{89}$$

and exponentiating the previous result, we obtain to first order

$$f_{\hat{\mathbf{P}}}(n) = \frac{\lambda^n}{n!}\mathrm{e}^{-\lambda} \xrightarrow{\lambda\to\infty} \frac{1}{\sqrt{2\pi\lambda}}\mathrm{e}^{-\frac{\delta^2}{2\lambda}} = \frac{1}{\sqrt{2\pi\lambda}}\mathrm{e}^{-\frac{(n-\lambda)^2}{2\lambda}} \tag{90}$$

5. Multivariate random variables

- Sometimes it is possible (and useful) to assign more that one random variable to the result of an experiment.
- For example, we can measure in a β-radioactive sample the time t and the speed v at which an electron is emitted; we can measure the time of arrival of the bus and the number of people in the waiting queue; we can observe whether it rains or not and measure the air temperature and pressure, and so on.
- In general, given an experiment, let us consider N random variables assigned to it: $(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N)$. The joint pdf of all these random variables is a function of N real variables $f_{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_N}(x_1, \ldots, x_N)$, which allows us to compute the probability that the vector of results $(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N)$ belongs to a region $\Omega \in \mathbb{R}^N$ as

$$P((\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N) \in \Omega) = \int_\Omega f_{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_N}(x_1, \ldots, x_N)\, dx_1 \ldots dx_N \tag{91}$$

- In other words, $f_{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_N}(x_1, \ldots, x_N)\, dx_1 \ldots dx_N$ is the probability that in a measurement of the N random variables the value of $\hat{\mathbf{x}}_1$ lies in $(x_1, x_1 + dx_1)$, the value of $\hat{\mathbf{x}}_2$ lies in $(x_2, x_2 + dx_2)$, and so on.
- The cumulative distribution function (cdf) is defined as

$$F_{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_N}(x_1, \ldots, x_N) = \int_{-\infty}^{x_1} dx'_1 \int_{-\infty}^{x_2} dx'_2 \ldots \int_{-\infty}^{x_N} dx'_N\, f_{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_N}(x'_1, x'_2, \ldots, x'_N) \tag{92}$$

- Statistical independence: N random variables $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N$ are defined to be statistically independent if the joint pdf factorizes as the product of pdfs for each variable, that is

$$f_{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_N}(x_1, \ldots, x_N) = f_{\hat{\mathbf{x}}_1}(x_1)\, f_{\hat{\mathbf{x}}_2}(x_2) \ldots f_{\hat{\mathbf{x}}_N}(x_N) \tag{93}$$

- The mean value of a function of N variables $G(x_1, ..., x_N)$ is computed as

$$\langle G(x_1, ..., x_N)\rangle = \int_{-\infty}^{\infty} dx_1 \ldots \int_{-\infty}^{\infty} dx_N \, G(x_1, ..., x_N) \, f_{\hat{\mathbf{x}}_1,...,\hat{\mathbf{x}}_N}(x_1, \ldots, x_N) \tag{94}$$

- In particular, if $G(x_1, ..., x_N) = \lambda_1 G_1(x_1) + \ldots + \lambda_N G_N(x_N)$, then

$$\langle G(x_1, ..., x_N)\rangle = \sum_{i=1}^{N} \lambda_i \langle G_i(x_i)\rangle \tag{95}$$

and if the random variables $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N$ are independent of each other, then

$$\sigma^2 \left[G(x_1, ..., x_N)\right] = \sum_{i=1}^{N} \lambda_i^2 \sigma^2 \left[G_i(x_i)\right] \tag{96}$$

- The covariance between two of the random variables $\hat{\mathbf{x}}_i$, $\hat{\mathbf{x}}_j$ is defined as

$$C[\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j] = C_{ij} = \langle (\hat{\mathbf{x}}_i - \langle \hat{\mathbf{x}}_i \rangle)(\hat{\mathbf{x}}_j - \langle \hat{\mathbf{x}}_j \rangle)\rangle = \langle \hat{\mathbf{x}}_i \hat{\mathbf{x}}_j \rangle - \langle \hat{\mathbf{x}}_i \rangle \langle \hat{\mathbf{x}}_j \rangle \tag{97}$$

Trivially, the covariance matrix is symmetrical, $C_{ij} = C_{ji}$. If the variables $\hat{\mathbf{x}}_i$, $\hat{\mathbf{x}}_j$ are statistically independent, then it is easy to verify that C_{ij} is also diagonal, i.e.

$$C_{ij} = \sigma^2[\hat{\mathbf{x}}_i]\delta_{ij} \tag{98}$$

where δ_{ij} is the Kronecker delta symbol. Note that the reverse statement (*"if $C_{ij} = \sigma^2[\hat{\mathbf{x}}_i]\delta_{ij}$ then variables $\hat{\mathbf{x}}_i$, $\hat{\mathbf{x}}_j$ are statistically independent"*) need not be true.

- In general, the variance of the sum of two functions $G_1(x)$, $G_2(x)$

$$\sigma^2[G_1 + G_2] = \langle (G_1 + G_2)^2 \rangle - \langle G_1 + G_2 \rangle^2 \,, \tag{99}$$

can be written as

$$\sigma^2[G_1 + G_2] = \sigma^2[G_1] + \sigma^2[G_2] + 2C[G_1, G_2] \tag{100}$$

where $C[G_1, G_2]$ is the covariance of G_1 and G_2.

- The correlation coefficient $\rho[\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j]$ of the random variables $\hat{\mathbf{x}}_i$, $\hat{\mathbf{x}}_j$ is defined as a suitable normalization of the covariance

$$\rho[\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j] = \frac{C[\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j]}{\sigma[\hat{\mathbf{x}}_i]\sigma[\hat{\mathbf{x}}_j]} \tag{101}$$

From the definition, it follows that $|\rho[\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j]| \leq 1$.

- Marginal distribution functions: Even if there are N random variables $(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N)$ defined in an experiment, we can still *"forget"* about some of them and consider the pdfs of only a subset of variables, for instance, $f_{\hat{\mathbf{x}}_1}(x_1)$ or $f_{\hat{\mathbf{x}}_2\hat{\mathbf{x}}_4}(x_2, x_4)$. These are called, in this context, *marginal* distribution functions and can be obtained by integrating out the variables that are not of interest. For example:

$$f_{\hat{\mathbf{x}}_1}(x_1) = \int_{-\infty}^{\infty} dx_2 \, f_{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}(x_1, x_2) \,, \tag{102}$$

or

$$f_{\hat{\mathbf{x}}_2\hat{\mathbf{x}}_4}(x_2, x_4) = \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_3 \, f_{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2\hat{\mathbf{x}}_3\hat{\mathbf{x}}_4}(x_1, x_2, x_3, x_4) \,, \tag{103}$$

6. Interpretation of the variance, statistical errors, and Chebyshev's theorem

- Let us consider a random variable $\hat{\mathbf{x}}$ assigned to an experiment. In general, every time we execute an experiment and obtain a result ξ, we do not know a priori which numerical value, $\hat{\mathbf{x}}(\xi)$, will the random variable take (unless there exists an event with probability 1). That is why it is called a random variable.

- Imagine we do know the average value $\mu = E[\hat{\mathbf{x}}]$ and the variance $\sigma^2 = E[\hat{\mathbf{x}}^2] - E[\hat{\mathbf{x}}]^2$. Maybe this knowledge comes from some theory that provides us with the values of μ and σ^2. What can we say about a single outcome $\hat{\mathbf{x}}(\xi)$ of the random variable? Not much in general.

- However we can say something about the probability of $\hat{\mathbf{x}}(\xi)$ taking values far away from μ, the mean value. Intuitively, we expect that it is unlikely to obtain values very far away from μ. But how unlikely?

- Chebyshev's theorem quantifies this probability as

$$P(|\hat{\mathbf{x}}(\xi) - \mu| \geq k\sigma) \leq \frac{1}{k^2}\,. \tag{104}$$

  for an arbitrary $k \geq 1$. In words, the probability that a single measurement $\hat{\mathbf{x}}(\xi)$ of a random variable differs from the mean value μ by an amount larger than k times the standard deviation σ is smaller than $1/k^2$.

- Proof of Chebyshev's inequality: Let $\hat{\mathbf{x}}$ be a random variable with pdf $f_{\hat{\mathbf{x}}}(x)$, with average μ and variance σ^2, and let $a \in \mathbb{R}^+$ be a positive number. Then

$$\langle \hat{\mathbf{x}}^2 \rangle = \int_{-\infty}^{\infty} x^2 f_{\hat{\mathbf{x}}}(x)\, dx \geq \int_{|x|\geq a} x^2 f_{\hat{\mathbf{x}}}(x)\, dx \geq a^2 \int_{|x|\geq a} f_{\hat{\mathbf{x}}}(x)\, dx = a^2 \mathrm{Prob}(|\hat{\mathbf{x}}| \geq a)\,. \tag{105}$$

or equivalently

$$\text{Prob}(|\hat{\mathbf{x}}| \geq a) \leq \frac{1}{a^2}\langle \hat{\mathbf{x}}^2 \rangle \,. \tag{106}$$

Applying this theorem to the random variable $\hat{\mathbf{y}} = \hat{\mathbf{x}} - \mu$, with zero average and the same variance σ^2, we obtain $P(|\hat{\mathbf{x}}(\xi) - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$, and taking the arbitrary constant $a = k\sigma$, we hence arrive at

$$P(|\hat{\mathbf{x}}(\xi) - \mu| \geq k\sigma) \leq \frac{1}{k^2} \,. \tag{107}$$

- This result can be written with the equivalent expression

$$P(|\hat{\mathbf{x}}(\xi) - \mu| \leq k\sigma) \geq 1 - \frac{1}{k^2} \,. \tag{108}$$

- Meaning: For instance, if $k = 3$, it is less than $1/3^2 \approx 11\%$ probable that the result of a single experiment lies outside the interval $(\mu - 3\sigma, \mu + 3\sigma)$. In other words, we cannot predict the result of a single experiment but we can affirm with an 11% confidence (about 89 out of every 100 times we make the experiment) that it will lie in the interval $(\mu - 3\sigma, \mu + 3\sigma)$. Of course, if σ is a large number, this prediction might be useless, but the reverse is also true, that is, if σ is small, then we can be pretty sure of the result.

- Example: Imagine that the experiment is to measure the radius of one polystyrene bead taken at random from a large number we bought from a manufacturer who tells us that the average radius of the set is $\mu = 3.5$ mm and the standard deviation is $\sigma = 1.0$ μm. How confident can we be that the radius of that particular bead lies in the interval (3.49, 3.51) mm? To apply Chebyshev's inequality to this data, we need to take $(3.49, 3.51) = (\mu - k\sigma, \mu + k\sigma)$ or $0.01\text{mm} = k \times 1\mu\text{m}$ or $k = 10$.This means that, on average, 1 out of $k^2 = 100$ beads will not have a radius within these limits (or, from the positive side, 99 out of 100 beads will have a radius within these limits).

- This interpretation of Chebyshev's theorem allows us to identify (in the precise manner defined before) the standard deviation of a distribution with the error (i.e. the uncertainty) in a single measurement of a random variable.

- Once we have understood this, we should understand the expression

$$\hat{\mathbf{x}}(\xi) = \mu \pm \sigma \tag{109}$$

with $\mu = E[\hat{\mathbf{x}}]$ and $\sigma^2 = E[\hat{\mathbf{x}}^2] - E[\hat{\mathbf{x}}]^2$ as a short-hand notation of the exact statement of Chebyshev's theorem (104). It does not mean that experimental values $\hat{\mathbf{x}}(\xi)$ that differ from μ by an amount greater than σ cannot appear, or are forbidden; it simply means that they are unlikely. How unlikely? Exactly by $1/k^2$, with $k = |\hat{\mathbf{x}}(\xi) - \mu|/\sigma$.

- Chebyshev's theorem is very general. It applies to any random variable whatever its pdf.

- In the case of a Gaussian distribution, we have the more precise result

$$P(|\hat{\mathbf{x}}(\xi) - \mu| \leq k\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mu-k\sigma}^{\mu+k\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \mathrm{erf}\left(\frac{k}{\sqrt{2}}\right) > 1 - \frac{1}{k^2}\,, \tag{110}$$

where we recall that the error function is defined as

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \mathrm{e}^{-y^2}\, dy,. \tag{111}$$

The previous probabilities take the following values for the Gaussian pdf

$$P(|\hat{\mathbf{x}}(\xi) - \mu| \leq \sigma) \;\; = 0.68269\ldots\,, \tag{112}$$
$$P(|\hat{\mathbf{x}}(\xi) - \mu| \leq 2\sigma) = 0.95450\ldots\,, \tag{113}$$
$$P(|\hat{\mathbf{x}}(\xi) - \mu| \leq 3\sigma) = 0.99736\ldots\,, \tag{114}$$

which means that we can be certain with a 68% probability that the result of the measurement will lie in the interval $(\mu - \sigma, \mu + \sigma)$; with a 95% probability that the result of the measurement will lie in the interval $(\mu - 2\sigma, \mu + 2\sigma)$; and with a 99.7% probability that the result of the measurement will lie in the interval $(\mu - 3\sigma, \mu + 3\sigma)$. Note that, if we take σ as the error of the measurement, in 32% (nearly 1/3) of the cases the observed value $\hat{\mathbf{x}}(\xi)$ will lie outside the error interval.

- In most cases, one does not know the distribution function of the experiment, neither the mean μ nor the standard deviation σ. Chebyshev's theorem can be read in the inverse sense

$$\mu = \hat{\mathbf{x}}(\xi) \pm \sigma \,. \tag{115}$$

  Given the result of a single measurement $\hat{\mathbf{x}}(\xi)$, this allows us to predict the value of μ within a certain interval of confidence which depends on the generally unknown standard deviation σ.

- However, it is clear that we cannot use this single measurement to obtain information about σ (which is ultimately related to the dispersion in a set of measurements). To obtain estimates for both μ and σ, we have to repeat the experiment n times, each one independent of the others, and use some properties of the sum of random variables.

7. Sum of random variables

- Let $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_n$ be independent random variables, all of them described by the same pdf $f_{\hat{\mathbf{x}}}(x)$ with mean μ and variance σ^2. The natural idea is to consider them as independent repetitions of the same experiment. Associated with the result $\Xi = (\xi_1, \xi_2, \ldots, \xi_n)$, we define the sample mean $\hat{\mathbf{S}}_n$ and the sample variance $\hat{\boldsymbol{\sigma}}_n^2$ as

$$\hat{\mathbf{S}}_n = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{x}}_i(\xi)\,, \tag{116}$$

$$\hat{\boldsymbol{\sigma}}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\mathbf{x}}_i(\xi) - \hat{\mathbf{S}}_n\right)^2\,,$$

$$= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{x}}_i(\xi)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{x}}_i(\xi)\right)^2\right)\,. \tag{117}$$

  The notation stresses the fact that both *random variables* depend on the number of repetitions n.

- It is easy to obtain the average of these two sample random variables[1] as

$$E[\hat{\mathbf{S}}_n] = E[\hat{\mathbf{x}}_i] = \mu\,, \tag{118}$$

$$E[\hat{\boldsymbol{\sigma}}_n^2] = \sigma^2\,. \tag{119}$$

  Furthermore, the variance of the sample mean is given by

$$\sigma^2[\hat{\mathbf{S}}_n] = E[\hat{\mathbf{S}}_n^2] - E[\hat{\mathbf{S}}_n]^2 = \frac{\sigma^2}{n}\,. \tag{120}$$

† The presence of the factor $n-1$ in Eq. (117) avoids its presence in Eq. (119).

- If we now repeat the experiment n times and obtain a value for $\hat{\mathbf{S}}_n(\Xi)$, we can use Chebyshev's theorem in its inverse short-hand notation (115) applied to the random variable $\hat{\mathbf{S}}_n$ to write $\mu = \hat{\mathbf{S}}_n \pm \sigma[\hat{\mathbf{S}}_n]$, or using Eq (120)

$$\mu = \hat{\mathbf{S}}_n(\Xi) \pm \frac{\sigma}{\sqrt{n}} \,. \tag{121}$$

- Still, we do not know the true value of σ on the right-hand side of this equation. It seems intuitive, though, given Eq. (119), that we can replace it by the sample variance $\sigma \approx \hat{\boldsymbol{\sigma}}_n$, leading to the final result

$$\mu = \hat{\mathbf{S}}_n(\Xi) \pm \frac{\hat{\boldsymbol{\sigma}}_n(\Xi)}{\sqrt{n}} \,. \tag{122}$$

  which yields an estimate of the average value together with its error. As discussed before, this error has to be interpreted in the statistical sense.

- As the sum of n independent random variables tends to a Gaussian distribution as n increases, see Section 9 on the Central Limit Theorem below, we can take the Gaussian confidence limits and conclude that in 68% of the cases the true value of μ will lie in the interval $(\hat{\mathbf{S}}_n(\Xi) - \frac{\hat{\boldsymbol{\sigma}}_n(\Xi)}{\sqrt{n}}, \hat{\mathbf{S}}_n(\Xi) + \frac{\hat{\boldsymbol{\sigma}}_n(\Xi)}{\sqrt{n}})$, and so on.

8. Law of large numbers

- The law of large numbers is a theorem that describes the result of performing the same experiment a large number of times. According to the law, the sample mean obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

- The law of large numbers is important because it "guarantees" stable long-term results for the averages of some random events. For example, while a casino may lose money in a single spin of the roulette wheel, its earnings will tend towards a predictable percentage over a large number of spins.

- Weak law of large numbers: Let $x_1, x_2, \ldots$ be an infinite sequence of i.i.d. Lebesgue integrable random variables with expected value $\langle x_i \rangle = \mu \; \forall i$. The weak law of large numbers (also called Khintchine's law) states that the sample average *converges in probability* towards the expected value

$$S_n = \frac{1}{n} \sum_{i=1}^{n} x_i \xrightarrow{n \to \infty} \mu \tag{123}$$

  This means in particular that, for any positive number ϵ,

$$\lim_{n \to \infty} \text{Prob}(|S_n - \mu| > \epsilon) = 0 \, . \tag{124}$$

- Interpreting this result, the weak law essentially states that for any nonzero margin specified, no matter how small, with a sufficiently large sample there will be a very high probability that the average of the observations will be close to the expected value; that is, within the margin.

- Convergence in probability is also called weak convergence of random variables. This version is called the weak law because random variables may converge weakly (in probability) as above without converging *strongly* (almost surely). Strong convergence means that $\text{Prob}(\lim_{n \to \infty} S_n = \mu) = 1$.

- The strong law implies the weak law but not vice versa: When the strong law conditions hold the variable converges both strongly (almost surely) and weakly (in probability). However the weak law may hold in conditions where the strong law does not hold and then the convergence is only weak (in probability).

- An assumption of finite variance $\sigma^2[x_1] = \sigma^2[x_2] = \ldots = \sigma^2 < \infty$ is not necessary. Large or infinite variance will make the convergence slower, but the law of large numbers holds anyway. This assumption is often used because it makes the proofs easier and shorter.

- Proof of the weak law of large numbers: We consider a sequence $x_1, x_2, \ldots$ of i.i.d. random variables with common pdf $f_{\hat{\mathbf{x}}}(x)$ with mean value μ and variance σ^2, and take the sample average $S_n = n^{-1}\sum_{i=1}^{n} x_i$. We now apply Chebyshev's inequality to the random variable $S_n - \mu$. Its expected value is $E[S_n - \mu] = 0$ and its variance is $\mathrm{Var}[S_n - \mu] = \sigma^2/n$, see central limit theorem below. Then Chebyshev's inequality implies that, for every $\epsilon > 0$,

$$\mathrm{Prob}(|S_n - \mu| \geq \epsilon) \leq \frac{1}{\epsilon^2}\mathrm{Var}(S_n - \mu) = \frac{1}{\epsilon^2}\frac{\sigma^2}{n}\,. \tag{125}$$

  Thus, for every $\epsilon > 0$, as $n \to \infty$ we find

$$\mathrm{Prob}(|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu| \geq \epsilon) \stackrel{n\to\infty}{=} 0 \tag{126}$$

- The law of large numbers can be also proven is simple terms using characteristic functions. The proof is morally equivalent to that of the central limit theorem below.

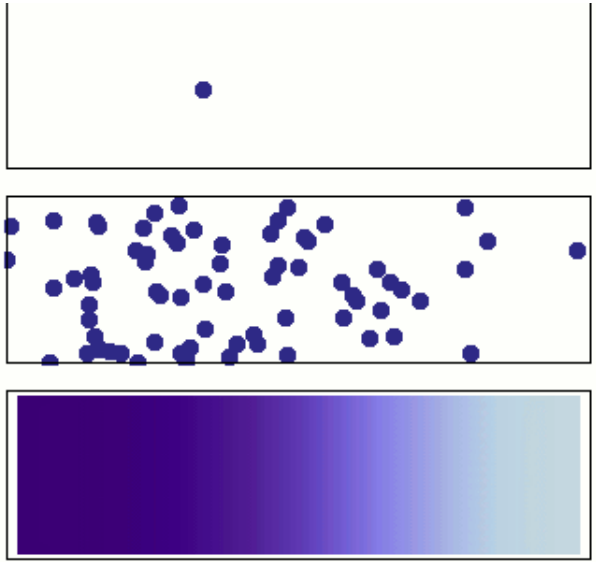

Figure 6. Diffusion and the law of large numbers

- Diffusion is an example of the law of large numbers, applied to chemistry. Initially, there are solute molecules on the left side of a barrier (purple line) and none on the right. The barrier is removed, and the solute diffuses to fill the whole container. Top: With a single molecule, the motion appears to be quite random. Middle: With more molecules, there is clearly a trend where the solute fills the container more and more uniformly, but there are also random fluctuations. Bottom: With an enormous number of solute molecules (too many to see), the randomness is essentially gone: The solute appears to move smoothly and systematically from high-concentration areas to low-concentration areas. In realistic situations, chemists can describe diffusion as a deterministic macroscopic phenomenon (see Fick's laws, $\partial_t \rho = \boldsymbol{\nabla} \cdot [D(\rho) \boldsymbol{\nabla} \rho]$), despite its underlying random nature.

9. Central limit theorem

- The central limit theorem (CLT) establishes that, for the most commonly studied scenarios, when independent random variables are added, their sum tends toward a normal distribution (commonly known as a bell curve) even if the original variables themselves are not normally distributed.

- In more precise terms, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined (finite) expected value and finite variance, will be approximately normally distributed, regardless of the underlying distribution.

- The theorem is a key concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions.

- Central limit theorem: Let $\{x_1, ..., x_n\}$ be a random sample of size n –that is, a sequence of n independent and identically distributed (i.i.d.) random variables drawn from an arbitrary distribution $f_{\hat{\mathbf{x}}}(x)$ of expected value given by μ and finite variance given by σ^2, and consider the sample average

$$S_n = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{127}$$

  Then, in the limit of large n, the pdf of the sample average S_n converges to a Gaussian (or normal) distribution of mean μ and variance σ^2/n, i.e.

$$f_{\hat{\mathbf{S}}_n}(S_n) \overset{n\to\infty}{=} \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left[-\frac{(S_n-\mu)^2}{2\sigma^2/n}\right] \tag{128}$$

- Proof: For a theorem of such fundamental importance to statistics and applied probability, the central limit theorem has a remarkably simple proof using characteristic functions. As stated above, suppose $\{x_1, ..., x_n\}$ to be a sequence of independent and identically distributed (i.i.d.) random variables drawn from an arbitrary distribution $f_{\hat{\mathbf{x}}}(x)$ of average μ and finite variance σ^2. To make the algebra simpler, consider now the excess sample average,

$$Z_n = S_n - \mu = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu) \equiv \frac{1}{n}\sum_{i=1}^{n} y_i \tag{129}$$

where we have defined the new random variable $\hat{\mathbf{y}} = \hat{\mathbf{x}} - \mu$, whose distribution $f_{\hat{\mathbf{y}}}(y)$ –of zero mean and variance σ^2– simply follows from $f_{\hat{\mathbf{x}}}(x)$ via conservation of probability. The pdf for Z_n can be simply written as

$$\begin{aligned} f_{\hat{\mathbf{Z}}_n}(Z_n) &= \int_{-\infty}^{\infty} dy_1 \ldots \int_{-\infty}^{\infty} dy_n \, f_{\hat{\mathbf{y}}_1,\ldots,\hat{\mathbf{y}}_n}(y_1, \ldots, y_N)\, \delta\left[Z_n - \frac{1}{n}\sum_{i=1}^{n} y_i\right] \\ &= \int_{-\infty}^{\infty} dy_1 \ldots \int_{-\infty}^{\infty} dy_n \, \prod_{i=1}^{n} f_{\hat{\mathbf{y}}_i}(y_i)\, \delta\left[Z_n - \frac{1}{n}\sum_{i=1}^{n} y_i\right], \end{aligned}$$

where the Dirac delta-function restricts the integrals to those n-tuples $(y_1, \ldots, y_n)$ whose sum is nZ_n. Note also that we have used the statistical independence of the different random variables in the second equality to write $f_{\hat{\mathbf{y}}_1,\ldots,\hat{\mathbf{y}}_n}(y_1, \ldots, y_N) = \prod_{i=1}^{n} f_{\hat{\mathbf{y}}_i}(y_i)$.

Working with a global constraint as the one imposed by the Dirac delta-function above is typically difficult, and taking now the characteristic function makes the problem considerably simpler by

eliminating the δ-function in favor of the exponential of the sum of random variables[1]. In particular,

$$G_{\hat{\mathbf{Z}}_n}(k) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}kZ_n}\, f_{\hat{\mathbf{Z}}_n}(Z_n)\, dZ_n = \int_{-\infty}^{\infty} dy_1 \ldots \int_{-\infty}^{\infty} dy_n\, \mathrm{e}^{\mathrm{i}\frac{k}{n}\sum_{l=1}^{n} y_l} \prod_{j=1}^{n} f_{\hat{\mathbf{y}}_j}(y_j)$$
$$= \left[\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\frac{k}{n}y}\, f_{\hat{\mathbf{y}}}(y)\, dy\right]^n = \left[G_{\hat{\mathbf{y}}}\left(\frac{k}{n}\right)\right]^n .$$

For a fixed value of k (not scaling with n), we have that $\frac{k}{n} \to 0$ as $n \to \infty$, so for pdf's with a finite variance we can always expand the characteristic function to first (non-zero) order as $G_{\hat{\mathbf{y}}}\left(\frac{k}{n}\right) \approx 1 - \frac{k^2}{2n^2}\sigma^2$, where we have used that $G'_{\hat{\mathbf{y}}}(0) = \mathrm{i}\mu = 0$ and $G''_{\hat{\mathbf{y}}}(0) = \mathrm{i}^2\langle\hat{\mathbf{y}}^2\rangle = -\sigma^2$. Therefore, using that $\lim_{n\to\infty}\left(1 - \frac{x}{n}\right)^n = \mathrm{e}^{-x}$, we find

$$G_{\hat{\mathbf{Z}}_n}(k) = \left[G_{\hat{\mathbf{y}}}\left(\frac{k}{n}\right)\right]^n \approx \left[1 - \frac{1}{n}\frac{k^2}{2n}\sigma^2\right]^n \xrightarrow{n\to\infty} \mathrm{e}^{-\frac{\sigma^2 k^2}{2n}} , \tag{130}$$

which is nothing but the characteristic function associated to a Gaussian pdf with zero mean and variance σ^2/n. Therefore

$$f_{\hat{\mathbf{Z}}_n}(Z_n) \overset{n\to\infty}{=} \frac{1}{\sqrt{2\pi\sigma^2/n}} \mathrm{e}^{-\frac{Z_n^2}{2\sigma^2/n}} \quad \Rightarrow \quad f_{\hat{\mathbf{S}}_n}(S_n) \overset{n\to\infty}{=} \frac{1}{\sqrt{2\pi\sigma^2/n}} \mathrm{e}^{-\frac{(S_n-\mu)^2}{2\sigma^2/n}} , \tag{131}$$

where we have reversed the initial change of variables, $S_n = Z_n + \mu$.

† This is a perfect example of the utility and convenience of using the characteristic function to solve a number of problems in statistics and probability.

9.1. Some examples and applications of the central limit theorem

- Since real-world quantities are often the balanced sum of many unobserved random events, the central limit theorem also provides a partial explanation for the prevalence of the normal probability distribution. It also justifies the approximation of large-sample statistics to the normal distribution in controlled experiments.

- A simple example is shown in Fig. 7, where the pdf of the sample mean of random 0s and 1s drawn from a binomial distribution is shown for different n's.

- A real-world example concerns the probability distribution for total distance covered in a random walk (biased or unbiased), which will tend toward a normal distribution.

- Flipping a large number of coins will result in a normal distribution for the total number of heads (or equivalently total number of tails). Indeed, if we assign a 1 for each head obtained, and a 0 for each tail, the the sum of this random variable will correspond to the total number of heads obtained, and hence its pdf according to the central limit theorem will obey a normal law.

- The central limit theorem explains the common appearance of the "Bell Curve" in density estimates applied to real world data. In cases like electronic noise, examination grades, and so on, we can often regard a single measured value as the weighted average of a large number of small effects. Using generalizations of the central limit theorem, we can then see that this would often (though not always) produce a final distribution that is approximately normal.

- Galton box, bean machine or quincunx: Invented by Sir Francis Galton to demonstrate the Central Limit Theorem, the machine consists of a vertical board with interleaved rows of pins. Balls are dropped from the top, and bounce either left or right as they hit the pins. Eventually, they are collected into

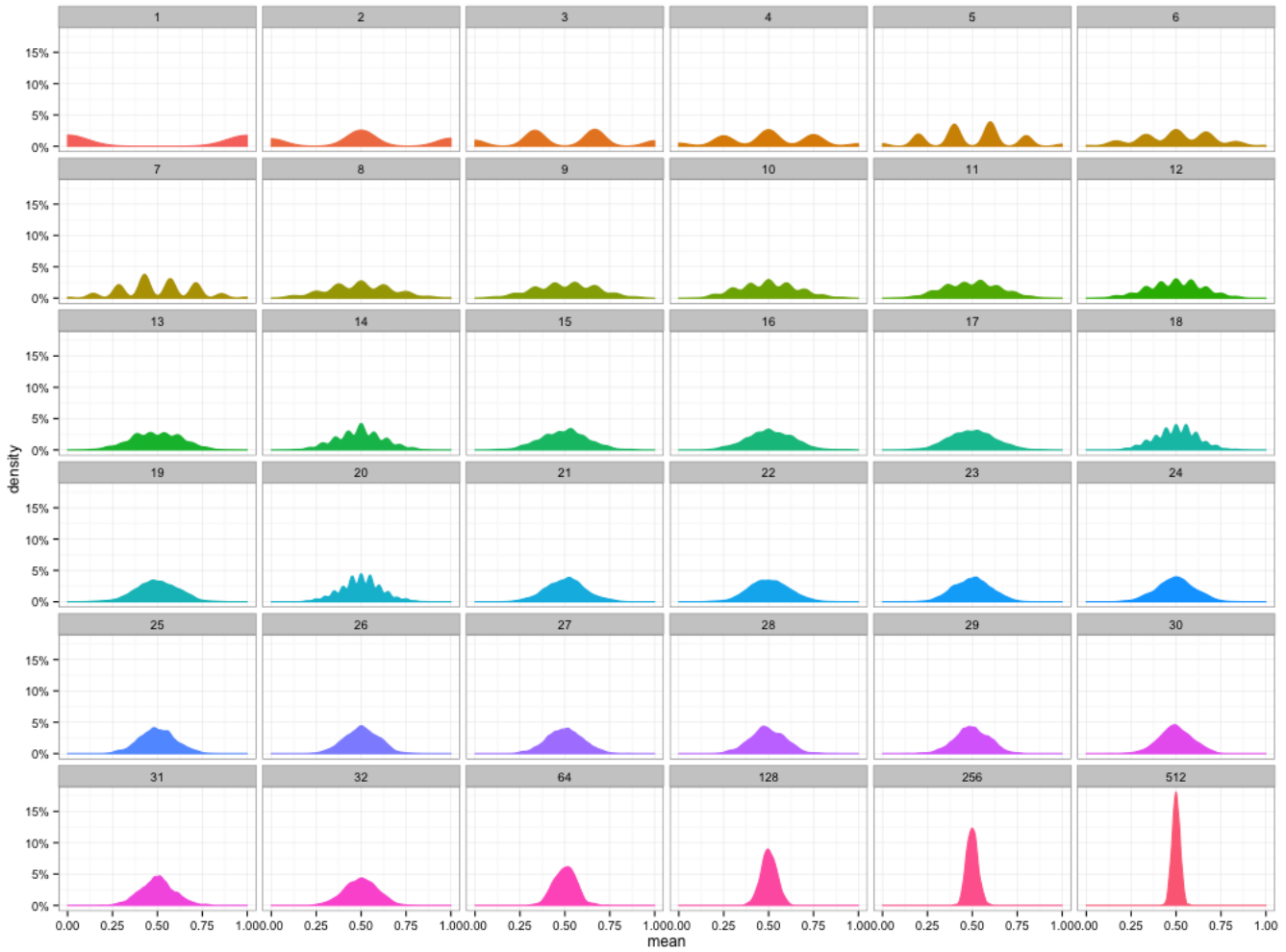

Figure 7. Pdf of the sample mean S_n for random variables drawn from the Bernuilli distribution. In particular, random 0s and 1s were generated, and then their means calculated for sample sizes ranging from $n = 1$ to $n = 512$. Note that as the sample size increases the tails become thinner and the distribution becomes more concentrated around the mean.

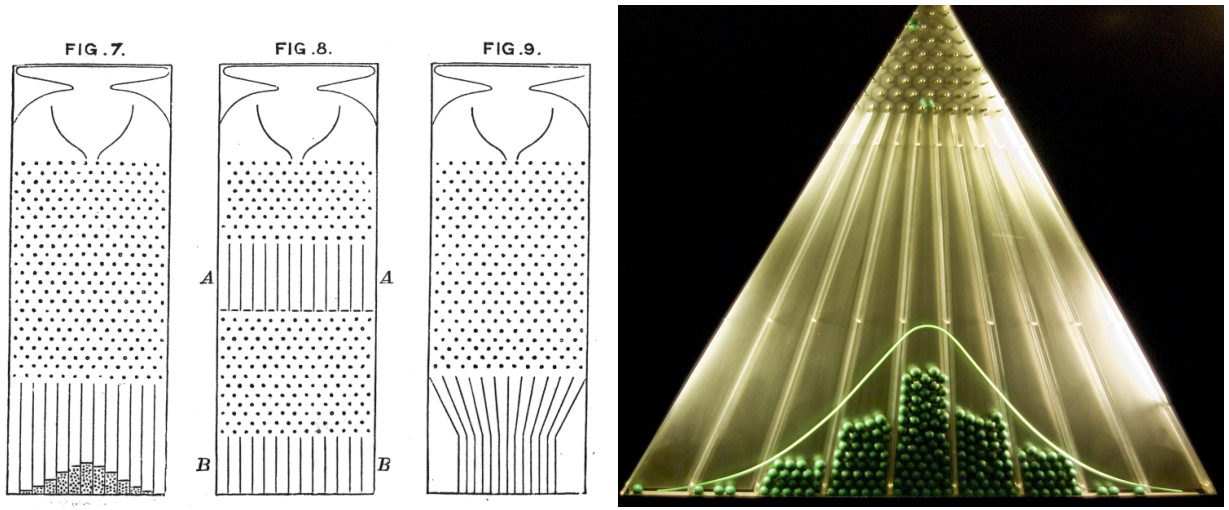

Figure 8. Left: The bean machine, as drawn by Sir Francis Galton. Right: A working replica of the machine (following a slightly modified design)

one-ball-wide bins at the bottom. The height of ball columns accumulated in the bins will eventually approximate a bell curve.

- Noise cancellation[1]: Suppose that a man is driving through the desert, and runs out of gas. He grabs his cellphone to make a call for help, dialing 911, but he is just at the edge of the broadcast range for his cellphone, and so his call to the 911 dispatcher is somewhat noisy and garbled. Suppose that the 911 dispatcher has the ability to use several cellphone towers to clean up the signal. Suppose that there are about 100 towers near to the stranded driver, and suppose that the signals they each receive at a particular instant in time is given by $X_1, \ldots, X_{100}$, where $X_i = S + \xi_i$, where S is the true signal being sent to the towers, and where ξ_i is the noise of each signal. Suppose that all the noises $\xi_1, \ldots, \xi_{100}$ are independent and identically distributed, and further suppose they all have mean 0 and variance σ^2. The

† Example taken from Ernie Croot's lecture notes at Georgia Tech

911 dispatcher cleans up the signal by computing the average

$$\bar{X} = \frac{1}{100}\sum_{i=1}^{100} X_i = S + \frac{1}{100}\sum_{i=1}^{100}\xi_i\,. \tag{132}$$

Now, by the Central Limit Theorem, we would expect that $\frac{1}{100}\sum_{i=1}^{100}\xi_i$ is distributed according to $N(0,\sigma^2/n)$, i.e. a normal, Gaussian function of mean 0 and variance $\sigma^2/100$. This yields 100-fold improvement in the noise variance gotten just using one tower! And hence the name *noise cancellation*.

10. Conditional probabilities

- For the sake of simplicity, we will consider the case of two random variables $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$, but similar ideas can be easily generalized in the case of more random variables.

- The joint probability density $f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)$ is defined such that the probability that a measurement of the random variable $\hat{\mathbf{x}}$ *and* the random variable $\hat{\mathbf{y}}$ gives for each one of them a value in the interval $(x, x+dx)$ and $(y, y+dy)$, respectively, is

$$P(x < \hat{\mathbf{x}} \leq x + dx, y < \hat{\mathbf{y}} \leq y + dy) = f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)\, dx\, dy \tag{133}$$

- The cumulative distribution function

$$F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(q, p)\, dq\, dp \tag{134}$$

  is such that

$$P(x_1 < \hat{\mathbf{x}} \leq x_2, y_1 < \hat{\mathbf{y}} \leq y_2) = F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x_2, y_2) - F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x_1, y_2) - F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x_2, y_1) + F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x_1, y_1) \tag{135}$$

- Some results follow straightforwardly from the definition:

$$\frac{\partial F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{\partial x} = \int_{-\infty}^{y} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, p)\, dp\,, \tag{136}$$

$$\frac{\partial F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{\partial y} = \int_{-\infty}^{x} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(q, y)\, dq\,, \tag{137}$$

$$\frac{\partial^2 F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{\partial x \partial y} = f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)\,. \tag{138}$$

- The marginal probabilities are

$$f_{\hat{\mathbf{x}}}(x) = \int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x,y)\,dy\,, \tag{139}$$

$$f_{\hat{\mathbf{y}}}(y) = \int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x,y)\,dx\,, \tag{140}$$

- Let us recall the definition of conditional probability. For any two events A and B such that $P(B) \neq 0$, the conditional probability of A given B is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{141}$$

- This suggests the definition of the conditional cumulative distribution function

$$F_{\hat{\mathbf{y}}}(y|B) = P(\hat{\mathbf{y}} \leq y|B) = \frac{P(\hat{\mathbf{y}} \leq y, B)}{P(B)}\,, \tag{142}$$

and the conditional density function

$$f_{\hat{\mathbf{y}}}(y|B) = \frac{\partial F_{\hat{\mathbf{y}}}(y|B)}{\partial y}\,. \tag{143}$$

- In the particular case of the event $B = \{\hat{\mathbf{x}} \leq x\}$, we have

$$F_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}} \leq x) = \frac{P(\hat{\mathbf{y}} \leq y, \hat{\mathbf{x}} \leq x)}{P(\hat{\mathbf{x}} \leq x)} = \frac{F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x,y)}{F_{\hat{\mathbf{x}}}(x)}\,, \tag{144}$$

and the pdf can be written as

$$f_{\hat{\mathbf{y}}}(y, \hat{\mathbf{x}} \leq x) = \frac{\partial F_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}} \leq x)}{\partial y} = \frac{\partial F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)/\partial y}{F_{\hat{\mathbf{x}}}(x)} = \frac{\displaystyle\int_{-\infty}^{x} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(q, y)\, dq}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{x} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(q, y)\, dq\, dy} \tag{145}$$

- If we now take $B = \{x_1 < \hat{\mathbf{x}} \leq x_2\}$, we get

$$F_{\hat{\mathbf{y}}}(y|x_1 < \hat{\mathbf{x}} \leq x_2) = \frac{P(x_1 < \hat{\mathbf{x}} \leq x_2, \hat{\mathbf{y}} \leq y)}{P(x_1 < \hat{\mathbf{x}} \leq x_2)} = \frac{F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x_2, y) - F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x_1, y)}{F_{\hat{\mathbf{x}}}(x_2) - F_{\hat{\mathbf{x}}}(x_1)} \tag{146}$$

and the pdf

$$f_{\hat{\mathbf{y}}}(y, x_1 < \hat{\mathbf{x}} \leq x_2) = \frac{\displaystyle\int_{x_1}^{x_2} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)\, dx}{\displaystyle\int_{x_1}^{x_2} f_{\hat{\mathbf{x}}}(x)\, dx} \tag{147}$$

- Let us consider, finally, the set $B = \{\hat{\mathbf{x}} = x\}$ as the limit $x_1 \to x_2$ of the previous case. Consequently, we define

$$F_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}} = x) = \lim_{\Delta x \to 0} F_{\hat{\mathbf{y}}}(y|x < \hat{\mathbf{x}} \leq x + \Delta x)\,. \tag{148}$$

From Eq. (146), we obtain

$$F_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}} = x) = \lim_{\Delta x \to 0} \frac{F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x + \Delta x, y) - F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{F_{\hat{\mathbf{x}}}(x + \Delta x) - F_{\hat{\mathbf{x}}}(x)} = \frac{\partial F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)/\partial x}{\partial F_{\hat{\mathbf{x}}}(x)/\partial x}\,, \tag{149}$$

which can be expressed as

$$F_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}} = x) = \frac{\int_{-\infty}^{y} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, p)\, dp}{f_{\hat{\mathbf{x}}}(x)}\,. \tag{150}$$

By takin derivative with respect to x, we obtain the conditional pdf

$$f_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}} = x) = \frac{f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{f_{\hat{\mathbf{x}}}(x)} = \frac{f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{\int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)\, dy}\,. \tag{151}$$

Moreover, exchanging the role of x and y, we obtain

$$F_{\hat{\mathbf{x}}}(x|\hat{\mathbf{y}} = y) = \frac{\int_{-\infty}^{x} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(q, y)\, dq}{f_{\hat{\mathbf{y}}}(y)}\,, \tag{152}$$

and

$$f_{\hat{\mathbf{x}}}(x|\hat{\mathbf{y}} = y) = \frac{f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{f_{\hat{\mathbf{y}}}(y)} = \frac{f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{\int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)\, dx}\,. \tag{153}$$

- For the sake of simplicity, and if no confusion can arise, we will shorten the notation of the four last defined functions to $F_{\hat{\mathbf{y}}}(y|x)$, $f_{\hat{\mathbf{y}}}(y|x)$, $F_{\hat{\mathbf{x}}}(x|y)$, and $f_{\hat{\mathbf{x}}}(x|y)$.
- By Bayes theorem, if A and B are events and B_1, B_2, ...is a partition of B, that is $B = \cup_i B_i$ and $B_i \cap B_j = \varnothing\ \forall i \neq j$, then

$$P(B_i|A) = \frac{P(A|B_i)\, P(B_i)}{\sum_j P(A|B_j)\, P(B_j)} \tag{154}$$

We now rephrase an equivalent Bayes theorem in terms of pdf'. It follows from Eqs. (151) and (153) that

$$f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x,y) = f_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}}=x)\, f_{\hat{\mathbf{x}}}(x)\,, \tag{155}$$
$$f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x,y) = f_{\hat{\mathbf{x}}}(x|\hat{\mathbf{y}}=y)\, f_{\hat{\mathbf{y}}}(y)\,, \tag{156}$$
$$\tag{157}$$

and therefore

$$f_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}}=x) = \frac{f_{\hat{\mathbf{x}}}(x|\hat{\mathbf{y}}=y)\, f_{\hat{\mathbf{y}}}(y)}{f_{\hat{\mathbf{x}}}(x)}\,. \tag{158}$$

We now use Eqs. (139) and (156) to derive

$$f_{\hat{\mathbf{x}}}(x) = \int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}}(x|\hat{\mathbf{y}}=y)\, f_{\hat{\mathbf{y}}}(y)\, dy \tag{159}$$

which when replaced in the denominator of Eq. (158) yields

$$f_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}}=x) = \frac{f_{\hat{\mathbf{x}}}(x|\hat{\mathbf{y}}=y)\, f_{\hat{\mathbf{y}}}(y)}{\int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}}(x|\hat{\mathbf{y}}=y)\, f_{\hat{\mathbf{y}}}(y)\, dy} \tag{160}$$

which is a version of Bayes theorem in terms of pdf's.

11. Markov chains

- It is not difficult to generalize these concepts of joint probabilities to more than two random variables. For example, the pdf of n random variables $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_n$ can be written in terms of conditional probabilities as

$$f_{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_n}(x_1,\ldots,x_n) = f_{\hat{\mathbf{x}}_1}(x_1)\, f_{\hat{\mathbf{x}}_2}(x_2|x_1)\, f_{\hat{\mathbf{x}}_3}(x_3|x_1,x_2)\ldots f_{\hat{\mathbf{x}}_n}(x_n|x_1,\ldots,x_{n-1}) \tag{161}$$

- This complicated expression adopts a much simpler form for a particular kind of random variables. A succession of random variables $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_n$ is called a Markov chain if for any value of $m = 1, \ldots, n$ it fulfills

$$f_{\hat{\mathbf{x}}_m}(x_m|x_1,\ldots,x_{m-1}) = f_{\hat{\mathbf{x}}_m}(x_m|x_{m-1})\,. \tag{162}$$

  That is, the pdf of $\hat{\mathbf{x}}_m$ conditioned to $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{m-1}$ is equal to the pdf of $\hat{\mathbf{x}}_m$ conditioned only to $\hat{\mathbf{x}}_{m-1}$. From this Markov property, Eq. (161) simplifies to

$$f_{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_n}(x_1,\ldots,x_n) = f_{\hat{\mathbf{x}}_n}(x_n|x_{n-1})\, f_{\hat{\mathbf{x}}_{n-1}}(x_{n-1}|x_{n-2})\ldots f_{\hat{\mathbf{x}}_2}(x_2|x_1)\, f_{\hat{\mathbf{x}}_1}(x_1) \tag{163}$$

- Therefore, the joint pdf of $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_n$ is determined only by the knowledge of $f_{\hat{\mathbf{x}}_1}(x_1)$ and the conditional pdfs $f_{\hat{\mathbf{x}}_m}(x|y)$ (also known in this context as the transition probability from y to x).

- A Markov chain is called *homogeneous* if the transition probabilities $f_{\hat{\mathbf{x}}_m}(x|y)$ are independent of m. Thus, for a homogeneous Markov chain, we write the transition probabilities simply as $f(x|y)$.

- It is easy to establish a relationship between $f_{\hat{\mathbf{x}}_{m+1}}(x)$ and $f_{\hat{\mathbf{x}}_m}(y)$ using the definition of conditional probability:

$$f_{\hat{\mathbf{x}}_{m+1}}(x) = \int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}_{m+1},\hat{\mathbf{x}}_m}(x,y)\,dy = \int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}_{m+1}}(x|y)\, f_{\hat{\mathbf{x}}_m}(y)\,dy \qquad m \geq 1\,, \tag{164}$$

which for a homogeneous chain reduces to

$$f_{\hat{\mathbf{x}}_{m+1}}(x) = \int_{-\infty}^{\infty} f(x|y)\, f_{\hat{\mathbf{x}}_m}(y)\, dy \quad m \geq 1\,. \tag{165}$$

- We can use this relation to construct the Markov chain. Starting from a given $f_{\hat{\mathbf{x}}_1}(x)$ initial pdf and a transition pdf $f(x|y)$, we can obtain the succession of random variables $\hat{\mathbf{x}}_m$, $m = 1, 2, \ldots$ with the respective pdfs $f_{\hat{\mathbf{x}}_m}(x)$.
- If the resulting pdfs $f_{\hat{\mathbf{x}}_m}(x)$ are all identical, $f_{\hat{\mathbf{x}}_m}(x) = f^{\rm st}_{\hat{\mathbf{x}}}(x)$, $m = 1, 2, \ldots$, we say that the Markov chain is stationary.
- Detailed balance: For a stationary Markov chain, Eq. (165) becomes

$$f^{\rm st}_{\hat{\mathbf{x}}}(x) = \int_{-\infty}^{\infty} f(x|y)\, f^{\rm st}_{\hat{\mathbf{x}}}(y)\, dy\,. \tag{166}$$

It is not easy, in general, to solve the above integral equation to find the stationary pdf of a Markov chain with a given transition pdf $f(x|y)$. However, using

$$\int_{-\infty}^{\infty} f(y|x)\, dy = \int_{-\infty}^{\infty} \frac{f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x,y)}{f_{\hat{\mathbf{x}}}(x)}\, dy = \frac{f_{\hat{\mathbf{x}}}(x)}{f_{\hat{\mathbf{x}}}(x)} = 1\,, \tag{167}$$

we can write Eq. (166) as

$$\begin{aligned} \int_{-\infty}^{\infty} f(y|x)\, f^{\rm st}_{\hat{\mathbf{x}}}(x)\, dy = & \int_{-\infty}^{\infty} f(x|y)\, f^{\rm st}_{\hat{\mathbf{x}}}(y)\, dy \Rightarrow \\ \Rightarrow & \int_{-\infty}^{\infty} \left[f(y|x)\, f^{\rm st}_{\hat{\mathbf{x}}}(x) - f(x|y)\, f^{\rm st}_{\hat{\mathbf{x}}}(y)\right] dy = 0\,. \end{aligned} \tag{168}$$

A way to satisfy this equation is by requiring the detailed balance condition

$$f(y|x)\, f^{\rm st}_{\hat{\mathbf{x}}}(x) = f(x|y)\, f^{\rm st}_{\hat{\mathbf{x}}}(y) \tag{169}$$

This is a simpler functional equation for $f^{\rm st}_{\hat{\mathbf{x}}}(x)$ than the integral Eq. (166). Any solution $f^{\rm st}_{\hat{\mathbf{x}}}(x)$ of the detailed balance condition will satisfy Eq. (166), but the reverse is not always true.

- Certainly, if a pdf $f^{\rm st}_{\hat{\mathbf{x}}}(x)$ satisfies Eq. (166), then it is a stationary solution of the recurrence relation (165) such that $f_{\hat{\mathbf{x}}_m}(x) = f^{\rm st}_{\hat{\mathbf{x}}}(x)$, $\forall m$, provided that $f_{\hat{\mathbf{x}}_1}(x) = f^{\rm st}_{\hat{\mathbf{x}}}(x)$.

- What happens when $f_{\hat{\mathbf{x}}_1}(x) \neq f^{\rm st}_{\hat{\mathbf{x}}}(x)$? Will the recurrence (165) converge toward the stationary solution $f^{\rm st}_{\hat{\mathbf{x}}}(x)$? A partial, but important, answer can be formulated as follows: If for every point x such that $f^{\rm st}_{\hat{\mathbf{x}}}(x) > 0$ and for every initial condition $f_{\hat{\mathbf{x}}_1}(x)$, there exists a number m of iterations such that $f_{\hat{\mathbf{x}}_m}(x) > 0$ (*irreducibility condition*) and the recurrence relation (165) does not get trapped in cyclic loops, then $f^{\rm st}_{\hat{\mathbf{x}}}(x)$ is the unique stationary solution and, furthermore, $\lim_{m\to\infty} f_{\hat{\mathbf{x}}_m}(x) = f^{\rm st}_{\hat{\mathbf{x}}}(x)$.

- These conditions (irreducibility and noncyclic behavior) are summarized by saying that the Markov chain is ergodic. The irreducibility condition has a simple intuitive interpretation. It states that, independent of the initial condition, the recurrence relation (165) does not have "forbidden" zones, meaning that it is able to provide eventually a pdf with a nonzero probability to any point x such that $f^{\rm st}_{\hat{\mathbf{x}}}(x) > 0$.

- Finally, we can consider that the variable m of the Markov chain represents, in some suitable units, a *time*. In this sense, Eq. (165) introduces a dynamics in the space of pdfs. We will often make use of this dynamical interpretation of a Markov chain in the rest of these notes.