Summary and problems: Probability and statistical thinking

Javier Aguilar

I. PROBABILITY SPACES: DISCRETE AND CONTINUOUS

I.1. For a continuous random variable, explain why $P(X = x) = 0$, yet $P(X \in [a, b])$ can be nonzero.

The probability that a continuous random variable lies in an interval is given by integrating its probability density function,

$$P(X \in [a, b]) = \int_a^b \rho(x)\, dx. \tag{1}$$

Hence,

$$P(X = a) = \int_a^a \rho(x)\, dx = 0. \tag{2}$$

I.2. Suppose X is discrete uniform over $\{1, \ldots, 10\}$. Compute $P(X \in \{3, 4, 8\})$.

Since these are disjoint events (the events cannot happen at the same time; the random variable is *either* 3, 4, 8, or some other value), the probabilities of such events are added:

$$P(X \in \{3, 4, 8\}) = 0.3. \tag{3}$$

I.3. Let X be continuous uniform over $[0, 1]$. Compute $P(X \in [0.2, 0.5])$.

$$P(X \in [0.2, 0.5]) = \int_{0.2}^{0.5} dx = 0.3. \tag{4}$$

I.4. Let X be a continuous variable with density $\rho(x) = 2x$ over $x \in [0, 1]$. Show that the distribution is normalized and compute $P(X > 0.5)$.

$$2\int_0^1 x\, dx = 1. \tag{5}$$

$$P(X > 0.5) = 2\int_{0.5}^1 x\, dx = 1 - 0.5^2 = \frac{3}{4}. \tag{6}$$

II. EXPECTED VALUES AND MOMENTS

Problems

II.1. What is the expected value of a constant? Compute $\langle\langle Y\rangle\rangle$.

Let $C \in \mathbb{R}$ be a constant, then by normalization of the probability density function,

$$\langle C\rangle = \int C\,\rho(x)\,dx = C. \tag{7}$$

By the exact reasoning,

$$\langle\langle Y\rangle\rangle = \langle Y\rangle. \tag{8}$$

II.2. Show that $var(X) = \langle X^2\rangle - \mu^2$.

$$var(X) = \langle (X - \langle X\rangle)^2\rangle = \langle X^2\rangle - 2\langle X\langle X\rangle\rangle + \langle X\rangle^2 = \langle X^2\rangle - \langle X\rangle^2. \tag{9}$$

II.3. Compute the mean and variance of a Bernuilli random variable $N \sim B(p)$ such that $P(N = 1) = p$ and $P(N = 0) = 1 - p$.

$$\langle X\rangle = \sum_n n\,P(X = n) = p. \tag{10}$$

$$\mathrm{var}(X) = \langle X^2\rangle - \langle X\rangle^2 = p - p^2 = p(1-p). \tag{11}$$

II.4. Compute the mean and variance of an exponential random variable $x \sim E(\alpha)$ such that $P(X \in [x, x + dx]) = \alpha\, e^{-\alpha x} dx$.

Integrals of the type $\int x^n e^{-x}$, with n an integer, can be solved by integration by parts, but also using parametric derivatives:

$$\langle X\rangle = \alpha \int_0^{+\infty} x e^{-\alpha x}\,dx = -\alpha\partial_\alpha \int_0^{+\infty} e^{-\alpha x} = -\alpha\partial_\alpha \frac{1}{\alpha} = \alpha^{-1}. \tag{12}$$

Similarly,

$$\langle X^2\rangle = \alpha \int_0^{+\infty} x^2 e^{-\alpha x}\,dx = \alpha\partial_\alpha^2 \int_0^{+\infty} e^{-\alpha x} = \alpha\partial_\alpha^2 \frac{1}{\alpha} = 2\alpha^{-2}. \tag{13}$$

Therefore,

$$\mathrm{var}(X) = \frac{1}{\alpha^2}. \tag{14}$$

II.5. Show that for any random variable X, $\langle I(X \in A)\rangle = P(X \in A)$.

Assuming that X is a continuous random variable taking values in a sampling space Ω,

$$P(X \in A) = \int_{x\in A} dx\, \rho(x) = \int_{x\in\Omega} dx\, \rho(x)\, I(X \in A) = \langle I(X \in A)\rangle. \tag{15}$$

II.6. Show that the power-law distribution is well-defined, but its moments may not exist. X is power law if $P(X \in [x, x + dx]) = \frac{\alpha-1}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-\alpha}$ for $x \geq x_{\min}$.

See discussion in Newman, M. E. J. "Power laws, Pareto distributions and Zipf's law" Contemporary Physics 46, 323–351 (2005).

III. JOINT AND CONDITIONED PROBABILITIES, AND BAYES THEOREM

III.1. What is $P(A|A)$ equal to?

The probability of A given that A is certain is one. Mathematically, this follows directly from the definition of conditional probability since $P(A, A) = P(A)$.

III.2. What is $P(A|B)$ if A and B are independent?

If A and B are independents, then $P(A, B) = P(A)P(B)$, and $P(A!B) = P(A)$. In other words, knowledge of B does not alter the probability of A if both events are independent.

III.3. In a fair die, consider $A = \{1, 2, 3\}$, $B = \{3, 6\}$. Compute $P(A)$, $P(B)$, $P(A, B)$, $P(A|B)$, and $P(B|A)$.

Since these are disjoint events,

$$P(A) = \frac{3}{6}, \tag{16}$$

and

$$P(B) = \frac{2}{6}. \tag{17}$$

The joint probability refers to the probability that both events occur simultaneously, which only happens if the result equals 3 3:

$$P(A, B) = \frac{1}{6}. \tag{18}$$

Therefore, the conditional probabilities are

$$P(A \mid B) = \frac{1}{2}, \tag{19}$$

and

$$P(B \mid A) = \frac{1}{3}. \tag{20}$$

III.4. Show that $P(A|B)$ is normalized as a function of A.

$$\sum_a P(A = a|B) = \frac{\sum_a P(A = a, B)}{P(B)} = 1, \tag{21}$$

Where we used marginalization.

III.5. Using the definition of conditioned probabilities, prove Bayes' theorem.

Joint probabilities can be conditioned with respect to all their arguments,

$$P(A, B) = P(A \mid B)P(B), \tag{22}$$

and also

$$P(A, B) = P(B \mid A)P(A). \tag{23}$$

Therefore,

$$P(A \mid B)P(B) = P(B \mid A)P(A), \tag{24}$$

from which Bayes' theorem follows directly.

IV. GALTON'S THEORY OF HUMAN HEIGHT HERITAGE

IV.1. Galton's model of height inheritance can be expressed using the figure of the "generant". The idea is that inheritance can happen in two ways: with probability p, a child inherits the height deviation (height minus population mean) of their parents; with probability $1 - p$, the child inherits the height of the "generant", whose height equals the population mean.

Assuming that the expected deviation of a child's height from the population mean is 2/3 of the parent's deviation. Under the model above, what should be the value of p?

Hint: Use the law of total expectation

$$\langle A \rangle = \sum_b \langle A \mid B = b \rangle \, P(B = b).$$

Here, let A be the child's inherited deviation, and let B be the event indicating whether height is inherited from the parents or from the generant. Compute:

- P(inherit from generant),
- P(inherit from parents),
- the average deviation given that height is inherited from the parents,
- the inherited deviation if height is inherited from the generant,

and use these to determine p.

Using the proposed formula with y_c denoting the deviation of sons and daughters and y_p the deviation of parents, let V_1 = height inherited from parents and V_2 = height inherited from the generant. Then

$$\langle y_c \rangle = \langle y_c \mid V_1 \rangle P(V_1) + \langle y_c \mid V_2 \rangle P(V_2) = \langle y_c \mid V_1 \rangle p + \langle y_c \mid V_2 \rangle (1 - p). \tag{25}$$

By definition, the deviation inherited from the generant is zero, and the deviation inherited from the parents equals the parental deviation. Therefore,

$$\langle y_c \rangle = p \, \langle y_p \rangle. \tag{26}$$

Therefore, using Galton's deviation ratio,

$$p = \frac{2}{3}. \tag{27}$$

IV.2. Prove Galton's claim: "it is more frequently the case that an exceptional man is the somewhat exceptional son of rather mediocre parents, than the average son of very exceptional parents", meaning that it is more likely that tall sons and daughters have parents with average heights.

Consider that both children and parents can either be average (A) or exceptional (E). Compute the probability that exceptional children have exceptional parents, and

the probability that exceptional children have average parents, using the following quantities:

$$P(C = E \mid P = E), \quad P(C = E \mid P = A), \quad P(P = A), \quad P(P = E).$$

What should be true about these probabilities for Galton's claim to hold?

Using the above reasoning, and the following empirical frequencies from Galton's data

$$P(h_P \leq 72 \text{ inches}) = 0.9, \tag{28}$$

$$P(h_C > 72 \text{ inches}) = 0.118, \tag{29}$$

$$P(h_C > 72.0 \text{ inches} | h_P > 72.0 \text{ inches}) = 0.556. \tag{30}$$

show that Galton's claim is correct.

We want to study the probability that exceptional (tall) children had tall parents, namely $P(P = E \mid C = E)$. Is this probability larger or smaller than $P(P = A \mid C = E)$? Both conditional probabilities share the same normalization factor and differ only in the joint probabilities in their numerators, which we now examine:

$$P(P = E \mid C = E) = \frac{P(C = E, P = E)}{P(C = E)}, \tag{31}$$

and

$$P(P = A \mid C = E) = \frac{P(C = E, P = A)}{P(C = E)}. \tag{32}$$

We can compute the joint probabilities using conditional probabilities, which may be easier to analyze:

$$P(C = E, P = E) = P(C = E \mid P = E)\, P(P = E). \tag{33}$$

Thus, even if exceptional parents are likely to have exceptional children ($P(C = E \mid P = E)$ should be high), the probability of being an exceptional parent is small ($P(P = E) \approx 0$). Therefore, the joint probability $P(C = E, P = E)$ may in principle be either small or moderately large depending on the balance of these factors.

A similar reasoning applies to the other joint probability,

$$P(C = E, P = A) = P(C = E \mid P = A)\, P(P = A), \tag{34}$$

since $P(C = E \mid P = A)$ is expected to be small, whereas $P(P = A)$ is high.

Using Galton's data and replacing $A \to h_p \leq 72$ inches, $E \to h_p > 72$ inches, we obtain

$$P(h_p > 72 \text{ inches} \mid h_c > 72 \text{ inches}) = \frac{P(h_c > 72 \text{ inches} \mid h_p > 72 \text{ inches}) P(h_p > 72 \text{ inches})}{P(h_c > 72 \text{ inches})} = 0.47. \tag{35}$$

Therefore, as Galton noted, it is more likely that tall sons and daughters have average parents. Although the difference is not large (both probabilities are close to 0.5), it is still remarkable that it is not true that most tall children have tall parents.

V. COMBINATORIAL, PROBABILITY COMPUTATION AND GOMBAUD'S BETTING PROBLEM.

V.1. Leibniz incorrectly argued that a sum of 11 and a sum of 12 are equally likely when throwing two dice. He claimed that "it is equally likely to throw twelve points as to throw eleven; because one or the other can be done in only one manner." What is wrong in Leibniz's reasoning?

It is true that there is only one combination of numbers giving twelve (6+6) and one giving eleven (6+5). However, the combination $5 + 6$ can occur in two different ways: $D_1 = 5$, $D_2 = 6$ and $D_1 = 6$, $D_2 = 5$, whereas the combination $6 + 6$ is compatible with only one outcome of the dice: $D_1 = 6$, $D_2 = 6$. Therefore, obtaining an eleven is twice as likely as obtaining a twelve.

This simple example illustrates the difference between unordered (or indistinguishable) statistics, in which we only care about the possible combinations of numbers (as Leibniz did), and ordered (or distinguishable) statistics, in which we recognize that the same combination of numbers can be produced in multiple ways.

V.2. Using the same strategy employed in Gombaud's problem, solve the birthday problem. What is the probability that, at least, two people in a class with n students share a birthday? In our class $n = 20$, what is the value of such probability? How many people are needed to reach a probability of 90% that at least two share a birthday? Produce a plot of probability vs group size.

We assume that birthdays are equally distributed along the year. The total number of possible birthday distributions is, according to the product rule,

$$N_B = 365^n. \tag{36}$$

Counting how coincidences of n birthdays are distributed is difficult; it is easier to count the number of ways in which there are no coincidences. This corresponds to the number of ways in which n people can select distinct days: the first person can select any of the 365 days, the second can select any day except the one selected by the first $(365-1)$, and so on, until the last person, who can select any of the $(365-n+1)$ remaining days. Therefore, the number of ways with no coincidence is

$$N_{nc} = \frac{365!}{(365-n)!}. \tag{37}$$

Therefore, the probability of no coincidence reads

$$p_{nc} = \frac{365!}{(365-n)!\,365^n}, \tag{38}$$

and the probability of, at least, one coincidence reads

$$p_1 = 1 - \frac{365!}{(365-n)!\,365^n}. \tag{39}$$

V.3. Newton–Pepys problem. Samuel Pepys asked Isaac Newton which of the following events has the highest probability:

A: At least one 6 appears when 6 fair dice are rolled.

B: At least two 6's appear when 12 fair dice are rolled.

C: At least three 6's appear when 18 fair dice are rolled.

As before, it is easier counting the number of ways in which the conditions is not held.

$$P(A) = 1 - P(\text{no 6 in 6 throws}) = 1 - \left(\frac{5}{6}\right)^6 = 0.66\ldots \tag{40}$$

$$\begin{aligned} P(B) &= 1 - P(\text{no 6 in 12 throws}) - P(\text{exactly one 6 in 12 throws}) \\ &= 1 - \left(\frac{5}{6}\right)^{12} - \frac{1}{6}\left(\frac{5}{6}\right)^{11}\binom{12}{1} = 0.61\ldots \end{aligned} \tag{41}$$

$P(B)$

$= 1 - P(\text{no 6 in 18 throws}) - P(\text{exactly one 6 in 18 throws}) - P(\text{exactly two 6 in 18 throws})$

$$= 1 - \left(\frac{5}{6}\right)^{18} - \frac{1}{6}\left(\frac{5}{6}\right)^{17}\binom{18}{1} - \left(\frac{1}{6}\right)^{2}\left(\frac{5}{6}\right)^{16}\binom{18}{2} = 0.59\ldots \tag{42}$$

Therefore, the game A is more likely to win.

VI. LAW OF LARGE NUMBERS, GAUSSIAN DISTRIBUTION AND CENTRAL LIMIT THEOREM

The law of large numbers states that if $\hat{\mu}$ is the empirical estimator of the mean of X measured from data,

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{43}$$

then this estimator converges to the expectation of the random variable as the number of samples grows:

$$\lim_{N\to\infty} \hat{\mu} = \langle X \rangle. \tag{44}$$

Let $X^{(i)}$, with $i = 1, \ldots, N$, be N independent and identically distributed random variables with $\mu = \langle X \rangle < \infty$ and $\sigma^2 = \mathrm{Var}(X) < \infty$. Then, if

$$Y = \frac{1}{N}\sum_{i=1}^{N} X^{(i)}, \tag{45}$$

the distribution of Y tends to a Gaussian as $N \to \infty$,

$$\rho(y)\,dy = P(Y \in [y, y+dy]) = \sqrt{\frac{N}{2\pi\sigma^2}}\,\exp\left[-N\frac{(y-\mu)^2}{2\sigma^2}\right]dy. \tag{46}$$

Problems

VI.1. Show that the mean and variance of the Gaussian

$$\rho(x)\,dx = P(X \in [x, x+dx]) = \frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

are, respectively, μ and σ^2. *Hint:* for the mean use symmetry; for the variance use the derivative of the normal integral,

$$\int_{-\infty}^{+\infty} dx\, e^{-a(x+b)^2} = \sqrt{\frac{\pi}{a}}.$$

We are given the Gaussian density

$$\rho(x)\,dx = P(X \in [x, x+dx]) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx.$$

Mean. The mean is

$$\langle X \rangle = \int_{-\infty}^{\infty} x\, \rho(x)\, dx.$$

With the change of variables $u = x - \mu$ (so $x = u + \mu$),

$$\langle X \rangle = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (u+\mu) e^{-u^2/(2\sigma^2)}\, du.$$

This becomes

$$\langle X \rangle = \frac{1}{\sqrt{2\pi\sigma^2}} \left(\underbrace{\int_{-\infty}^{\infty} u\, e^{-u^2/(2\sigma^2)}\, du}_{0} + \mu \int_{-\infty}^{\infty} e^{-u^2/(2\sigma^2)}\, du \right),$$

where the first integral vanishes by symmetry. Using the normalization of the Gaussian integral,

$$\langle X \rangle = \mu.$$

Variance. The variance is

$$\langle (X-\mu)^2 \rangle = \int_{-\infty}^{\infty} (x-\mu)^2\, \rho(x)\, dx.$$

Again with $u = x - \mu$,

$$\langle (X-\mu)^2 \rangle = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} u^2 e^{-u^2/(2\sigma^2)}\, du.$$

Using the Gaussian integral identity

$$\int_{-\infty}^{+\infty} e^{-a(x+b)^2}\, dx = \sqrt{\frac{\pi}{a}},$$

differentiate with respect to a (and set $b = 0$) to get

$$\int_{-\infty}^{\infty} x^2 e^{-ax^2}\, dx = \frac{1}{2}\sqrt{\pi}\, a^{-3/2}.$$

Setting $a = \frac{1}{2\sigma^2}$ yields

$$\int_{-\infty}^{\infty} u^2 e^{-u^2/(2\sigma^2)}\, du = \sqrt{2}\,\sqrt{\pi}\,\sigma^3.$$

Thus,

$$\langle (X-\mu)^2 \rangle = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \sqrt{2}\,\sqrt{\pi}\,\sigma^3 = \sigma^2.$$

Therefore the Gaussian has mean $\langle X \rangle = \mu$ and variance $\langle (X-\mu)^2 \rangle = \sigma^2$.

VI.2. Show that the central limit theorem applies to the empirical estimator of the mean, $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$. Show explicitly that the central limit theorem implies the law of large numbers.

Since the sample mean is a sum of identically distributed random variables, the central limit theorem applies. In particular, the theorem states that $\hat{\mu}$ tends to a Gaussian distribution with mean $\mu = \langle X \rangle$ and variance

$$\mathrm{var}(\hat{\mu}) = \frac{\mathrm{var}(X)}{N}. \tag{47}$$

In the limit $N \to \infty$, the variance vanish and the Gaussian distribution tends to a Dirac delta,

$$\lim_{N\to\infty} P\left(\hat{\mu} \in [u, u+du]\right) = \delta\left(u - \langle X \rangle\right). \tag{48}$$

Therefore, as N tends to infinity, the value of $\hat{\mu}$ (not its moment) tend to $\langle X \rangle$, which is the statement of the law of large numbers.

VI.3. The 68–95–99.7 rule: Let X be Gaussian with mean μ and variance σ^2. Compute

$$P(X \in [\mu-\sigma, \mu+\sigma]), \quad P(X \in [\mu-2\sigma, \mu+2\sigma]), \quad P(X \in [\mu-3\sigma, \mu+3\sigma]).$$

Hint: use numerical integration or the error function.

Let $Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$. By symmetry,

$$P(X \in [\mu-k\sigma, \mu+k\sigma]) = P(|Z| \le k) = \int_{-k}^{k} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz.$$

Using the error function $\mathrm{erf}(x) = \dfrac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$ we get

$$P(|Z| \le k) = \mathrm{erf}\left(\frac{k}{\sqrt{2}}\right).$$

Hence

$$P(X \in [\mu-\sigma, \mu+\sigma]) = \text{erf}\left(\frac{1}{\sqrt{2}}\right) \approx 0.682689 \quad (68.27\%),$$

$$P(X \in [\mu-2\sigma, \mu+2\sigma]) = \text{erf}\left(\frac{2}{\sqrt{2}}\right) = \text{erf}(\sqrt{2}) \approx 0.954500 \quad (95.45\%),$$

$$P(X \in [\mu-3\sigma, \mu+3\sigma]) = \text{erf}\left(\frac{3}{\sqrt{2}}\right) \approx 0.997300 \quad (99.73\%).$$

VI.4. A sum of Bernuilli random variables is a binomial random variable. Compute explicitly how normalized sums of binomial random variables converge to Gaussian distributions. *Hint:* $X \sim \text{Binomial}(N, p)$. First show that in the limit $N \to \infty$, $p \to 0$, with $\lambda = Np$ fixed, the binomial distribution converges to a Poisson distribution. Then show that for large λ the Poisson converges to a Gaussian, using Stirling's approximation. **From Binomial to Poisson:**

$$\begin{aligned} P(X=k) &= p^k(1-p)^{N-k}\frac{N!}{k!(N-k)!} = \frac{p^k(1-p)^{N-k}}{k!}N(N-1)\dots(N-k+1) \\ &= \frac{(Np)^k(1-p)^{N-k}}{k!} + \mathcal{O}\left(\frac{(Np)^k}{N}\right). \end{aligned} \tag{49}$$

We now take the limit $p \to 0$ and $N \to \infty$ with $Np = \lambda$ held constant. Using

$$\lim_{\substack{p\to 0 \\ N\to\infty}} (1-p)^{N-k} = \lim_{\substack{p\to 0 \\ N\to\infty}} e^{-Np+kp} = e^{-\lambda}, \tag{50}$$

we obtain

$$\lim_{\substack{p\to 0 \\ N\to\infty}} P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{51}$$

From Poisson to Gaussian:

Let X be a Poisson random variable with parameter λ,

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{52}$$

Define a new random variable that has finite mean and variance as $\lambda \to \infty$,

$$Y = \frac{X-\lambda}{\sqrt{\lambda}}. \tag{53}$$

The distribution of Y is

$$\rho(y) = \frac{\lambda^{k(y)}e^{-\lambda}}{k(y)!}\sqrt{\lambda}, \tag{54}$$

with $k(y) = \sqrt{\lambda}\, y + \lambda$.

Using Stirling–de Moivre's approximation,

$$n! \sim \sqrt{2\pi n}\, n^n e^{-n}, \tag{55}$$

we find

$$\rho(y) \approx \frac{e^{\sqrt{\lambda} y}}{\sqrt{2\pi}} \left(\frac{k(y)}{\lambda} \right)^{-k(y)} = \frac{e^{\sqrt{\lambda} y}}{\sqrt{2\pi}} \left(1 + \frac{y}{\sqrt{\lambda}} \right)^{-\sqrt{\lambda} y - \lambda}. \tag{56}$$

If one naively uses the limit

$$1 + \frac{y}{\sqrt{\lambda}} = e^{y/\sqrt{\lambda}},$$

then an essential correction term is lost. Instead we expand the logarithm:

$$\log\left(1 + \frac{y}{\sqrt{\lambda}} \right) = \frac{y}{\sqrt{\lambda}} - \frac{y^2}{2\lambda} + \mathcal{O}(\lambda^{-3/2}). \tag{57}$$

$$\begin{aligned} \rho(y) &\approx \frac{e^{\sqrt{\lambda} y}}{\sqrt{2\pi}} \exp\left[-(\sqrt{\lambda} y + \lambda) \left(\frac{y}{\sqrt{\lambda}} - \frac{y^2}{2\lambda} + \mathcal{O}(\lambda^{-3/2}) \right) \right] \\ &= \frac{e^{-y^2/2 + \mathcal{O}(\lambda^{-1/2})}}{\sqrt{2\pi}}. \end{aligned} \tag{58}$$

Therefore, in the limit $\lambda \to \infty$, Y converges to a Gaussian random variable with zero mean and unit variance.

VII. THE MONTY HALL PROBLEM

VII.1. Complete the calculation of winning probabilities in the Monty Hall problem using conditional probabilities.

Let us consider two random variables: C is the door where the car is, which can be any of the three doors, $C \in \{1, 2, 3\}$. M is the door opened by Monty Hall, which can also be $M \in \{1, 2, 3\}$. By symmetry, the computation of probabilities is independent of the label of the door that is initially chosen by the player. Therefore, we assume without loss of generality that the initially chosen door is door 1. Also without loss of generality, let us assume that $M = 2$, so that Monty opens the second door and reveals no prize. The aim is to compute the probability of winning by staying with the first

door (the "no change strategy") and by switching to the third door (the "changing strategy"). In mathematical terms, we want to compute

$$P(\text{Win with no change}) = P(C = 1 \mid M = 2), \tag{59}$$

and

$$P(\text{Win changing}) = P(C = 3 \mid M = 2). \tag{60}$$

By the rules of the game, it is impossible that the prize is behind the door Monty opens. Therefore $P(C = 2 \mid M = 2) = 0$, and

$$P(C = 3 \mid M = 2) = 1 - P(C = 1 \mid M = 2).$$

We use Bayes' theorem to compute the target probability:

$$P(C = 1|M = 2) = \frac{P(M = 2|C = 1)P(C = 1)}{P(M = 2)} \tag{61}$$

By symmetry, $P(C = 1) = 1/3$. If the car is in the first door, Monty has no preferences to show the second or third door, and therefore $P(M = 2|C = 1) = 0.5$. The probability $P(M = 2)$ is computed by marginalization,

$$P(M = 2) = \sum_{d=1,2,3} P(M = 2|C = d)\, P(C = d) = \frac{1}{3}\left[P(M = 2|C = 1) + P(M = 2|C = 3)\right]. \tag{62}$$

If the car is in the third door, but we chose the first door, Monty is forced to show what is behind the second door (Monty cannot show the prize). Therefore, $P(M = 2|C = 3) = 1$. Substituting all the computed probabilities we find

$$P(C = 1|M = 2) = 1/3. \tag{63}$$

Therefore, $P(C = 3|M = 2) = 2/3$, and is more likely to win the prize if we change the door after Monty has given his clue.

VII.2. Repeat the calculation of winning probabilities in the Monty Hall problem using a decision tree (i.e. by enumerating all possible outcomes of the game and computing the probabilities as the number of favorable cases divided by the total number of possible games).

As before, by symmetry we assume the player initially chooses door 1. Let $C \in \{1,2,3\}$ be the door with the car and $M \in \{1,2,3\}$ the door Monty opens. Monty never opens the door with the car and if he has a choice between two goat doors he picks uniformly at random.

We build the decision tree by first branching on the car location (each with probability 1/3), then on Monty's choice:

- $C = 1$ (probability 1/3). Monty has a choice and opens door 2 or door 3 with equal probability:

$$(C = 1, M = 2) \quad \text{with probability} \quad \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6},$$
$$(C = 1, M = 3) \quad \text{with probability} \quad \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}.$$

- $C = 2$ (probability 1/3). Monty cannot open door 2, so he must open door 3:

$$(C = 2, M = 3) \quad \text{with probability} \quad \frac{1}{3} \cdot 1 = \frac{1}{3}.$$

- $C = 3$ (probability 1/3). Monty must open door 2:

$$(C = 3, M = 2) \quad \text{with probability} \quad \frac{1}{3} \cdot 1 = \frac{1}{3}.$$

Thus the (distinct) outcome-probabilities are

$$(C = 1, M = 2) : \tfrac{1}{6}, \qquad (C = 1, M = 3) : \tfrac{1}{6}, \qquad (C = 2, M = 3) : \tfrac{1}{3}, \qquad (C = 3, M = 2) : \tfrac{1}{3}.$$

To perform counting ("number of favorable cases / total number of possible games") convert probabilities to integer counts by multiplying by the common denominator 6:

$$(C = 1, M = 2) : 1, \qquad (C = 1, M = 3) : 1, \qquad (C = 2, M = 3) : 2, \qquad (C = 3, M = 2) : 2.$$

These six count-units represent the full sample space (they are proportional to the outcome probabilities).

Now count favorable cases.

a. No change (stay with door 1). You win iff the car is behind door 1, i.e. outcomes $(C = 1, M = 2)$ or $(C = 1, M = 3)$. Count of favorable units: $1 + 1 = 2$. Total units: 6. Hence

$$P(\text{win by staying}) = \frac{2}{6} = \frac{1}{3}.$$

b. Change (switch to the other unopened door). If Monty opened door 2 you switch to door 3 and win exactly when $C = 3$ (outcome $(C = 3, M = 2)$, count 2). If Monty opened door 3 you switch to door 2 and win exactly when $C = 2$ (outcome $(C = 2, M = 3)$, count 2). Total favorable units: $2 + 2 = 4$. Therefore

$$P(\text{win by switching}) = \frac{4}{6} = \frac{2}{3}.$$

c. Conclusion. Counting the decision-tree outcomes shows staying wins in 2 of 6 equally weighted count-units (1/3) while switching wins in 4 of 6 (2/3). Thus switching doubles your probability of winning.

VIII. INFERENCE: PARAMETER ESTIMATION OF PROBABILITY DISTRIBUTIONS

VIII.1. Using both the method of moments and maximum likelihood estimation, derive the parameter estimators for the Bernoulli, Poisson, binomial, Gaussian (mean and variance), exponential, and power law distributions. For each case, compute the variance of the estimator and discuss the differences between the two methods.

Bernoulli

$$P(X = n) = \begin{cases} p, & \text{if } n = 1, \\ 1 - p, & \text{if } n = 0. \end{cases} \tag{64}$$

Since

$$\langle X \rangle = p, \tag{65}$$

then

$$\hat{p}_{ME} = \frac{1}{M} \sum_{i=1}^{M} x_i. \tag{66}$$

The likelihood reads

$$L = p^n \left(1 - p\right)^{M-n}, \tag{67}$$

where $n = \sum_{i=1}^{M} x_i$.

$$\partial_p L = p^n \left(1 - p\right)^{M-n} \left(\frac{n}{p} - \frac{M - n}{1 - p} \right). \tag{68}$$

Therefore, for $p \in (0, 1)$,

$$\partial_p L = 0 \iff n(1 - p) - (M - n)p = 0. \tag{69}$$

Therefore,

$$\hat{p}_{ME} = \hat{p}_{MLE} = \frac{n}{M}. \tag{70}$$

Now, assuming a uniform prior for p and treating p as a random variable, the posterior density reads

$$\rho(p) = Z^{-1} p^n \left(1-p\right)^{M-n}, \tag{71}$$

with

$$Z = \int_0^1 dp\, p^n \left(1-p\right)^{M-n} = B(n+1, M-n+1), \tag{72}$$

where $B(\cdot,\cdot)$ is the beta function,

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{73}$$

The distribution for p is therefore the beta distribution, whose summary statistics are:

$$E_B(p) = \frac{n+1}{M+2}, \tag{74}$$

$$\text{mode}_B(p) = \frac{n}{M}, \tag{75}$$

$$\text{var}_B(p) = \frac{(n+1)(M-n+1)}{(M+2)^2(M+3)}. \tag{76}$$

Exponential distribution

$$P(X \in [x, x+dx]) = \lambda e^{-\lambda x} dx. \tag{77}$$

Since

$$\langle X \rangle = \frac{1}{\lambda}, \tag{78}$$

then

$$\hat{\lambda}_{ME} = \frac{1}{\hat{\mu}}, \tag{79}$$

with

$$\hat{\mu} = \frac{1}{M} \sum_{i=1}^{M} x_i. \tag{80}$$

The likelihood reads

$$L = \lambda^M e^{-\lambda M \hat{\mu}}, \tag{81}$$

and its derivative

$$\partial_\lambda L = L \left(\frac{M}{\lambda} - M\hat{\mu} \right). \tag{82}$$

Therefore,

$$\hat{\lambda}_{MLE} = \hat{\lambda}_{ME} = \frac{1}{\hat{\mu}}. \tag{83}$$

Now let us consider λ as a random variable and study its distribution if an uniform prior is assumed,

$$\rho(\lambda) = \lambda^M e^{-\lambda M\hat{\mu}} Z^{-1}, \tag{84}$$

with

$$Z = \int_0^\infty \lambda^M e^{-\lambda M\hat{\mu}}\, d\lambda = \frac{\Gamma(M+1)}{(M\hat{\mu})^{M+1}}. \tag{85}$$

Therefore, λ is gamma-distributed and its summary statistics are known to be:

$$\langle\lambda\rangle = \frac{M+1}{M\hat{\mu}}, \tag{86}$$

$$\text{mode}(\lambda) = \frac{1}{\hat{\mu}}, \tag{87}$$

$$\text{var}(\lambda) = \frac{M+1}{(M\hat{\mu})^2}. \tag{88}$$

Poisson distribution

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

Since

$$\langle X\rangle = \lambda, \tag{90}$$

then

$$\hat{\lambda}_{ME} = \frac{1}{M}\sum_{i=1}^{M} x_i. \tag{91}$$

The likelihood reads

$$L = \frac{e^{-M\lambda}\lambda^\alpha}{\beta}, \tag{92}$$

where $\alpha = \sum_{i=1}^M x_i$, and $\beta = \prod_{i=1}^M x_i!$.

$$\partial_\lambda L = L\left(\frac{\alpha}{\lambda} - M\right), \tag{93}$$

Since $L > 0$ for finite number of measures,

$$\hat{\lambda}_{ME} = \hat{\lambda}_{MLE} = \frac{\alpha}{M}. \tag{94}$$

Now, assuming a uniform prior for λ and treating λ as a random variable, the posterior density reads

$$\rho(\lambda) = e^{-M\lambda}\lambda^\alpha Z^{-1}, \tag{95}$$

with

$$Z = \int_0^\infty d\lambda\, e^{-M\lambda}\lambda^\alpha = \frac{\Gamma(\alpha+1)}{M^{\alpha+1}}. \tag{96}$$

The distribution for λ is therefore the gamma distribution, whose summary statistics are:

$$E_G(\lambda) = \frac{\alpha+1}{M}, \tag{97}$$

$$\mathrm{mode}_G(\lambda) = \frac{\alpha}{M}, \tag{98}$$

$$\mathrm{var}_G(\lambda) = \frac{\alpha+1}{M^2}. \tag{99}$$

Binomial

$$P(X=n) = p^n(1-p)^{M-n}\binom{M}{n}. \tag{100}$$

The centered moments of the binomial read

$$\langle X\rangle = Mp, \tag{101}$$

and

$$\mathrm{Var}(X) = Mp(1-p), \tag{102}$$

Therefore,

$$p = 1 - \frac{\mathrm{Var}(X)}{\langle X\rangle}, \tag{103}$$

With this formula, we build the moment estimators as

$$\hat{p}_{ME} = 1 - \frac{\hat{\sigma}^2}{\hat{\mu}}, \tag{104}$$

and

$$\hat{M}_{ME} = \frac{\hat{\mu}}{\hat{p}_{ME}}, \tag{105}$$

with

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^m x_i, \quad \hat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^m x_i^2 - \hat{\mu}^2. \tag{106}$$

Now let us look for the MLE estimators,

$$L = p^{m\hat{\mu}}(1-p)^{m(M-\hat{\mu})}\prod_{i=1}^m \binom{M}{n_i}. \tag{107}$$

It might be easier to operate with the log-likelihood

$$\ell = \log(L) = m\,\hat{\mu}\log(p) + m(M-\hat{\mu})\log(1-p) + \sum_{i=1}^m \log\left[\binom{M}{n_i}\right]. \tag{108}$$

There is no problem on deriving the log-likelihood with respect to p,

$$\partial_p \ell = \frac{m\hat{\mu}}{p} - \frac{m(M-\hat{\mu})}{1-p}. \tag{109}$$

For fixed M, we could find the MLE,

$$\hat{p}_{MLE} = \frac{\hat{\mu}}{M}. \tag{110}$$

However, since M is discrete, we cannot maximize the likelihood in the $M-$direction by derivation. One should replace p by its MLE estimation and maximize numerically giving values to M (grid search). One could work-out approximations using Stirling's formula, but still there is no analytical MLE equation for M.

Gaussian distribution Gaussian distribution

$$P(X \in [x, x+dx]) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx. \tag{111}$$

Since

$$\langle X \rangle = \mu, \qquad \mathrm{Var}(X) = \sigma^2, \tag{112}$$

the method-of-moments estimators are

$$\hat{\mu}_{ME} = \hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i, \tag{113}$$

$$\hat{\sigma}^2_{ME} = \frac{1}{M}\sum_{i=1}^{M} (x_i - \hat{\mu})^2. \tag{114}$$

The likelihood for an i.i.d. sample $\{x_i\}_{i=1}^M$ is

$$L(\mu, \sigma^2) = (2\pi\sigma^2)^{-M/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{M}(x_i-\mu)^2\right]. \tag{115}$$

Differentiating w.r.t. μ and equating to zero gives

$$\partial_\mu \log L = \frac{1}{\sigma^2}\sum_{i=1}^{M}(x_i-\mu) = 0 \quad \Longrightarrow \quad \hat{\mu}_{MLE} = \frac{1}{M}\sum_{i=1}^{M} x_i = \hat{\mu}. \tag{116}$$

Differentiating w.r.t. σ^2 and equating to zero gives

$$\partial_{\sigma^2} \log L = -\frac{M}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{M}(x_i-\mu)^2 = 0, \tag{117}$$

so, substituting $\mu = \hat{\mu}_{MLE}$,

$$\hat{\sigma}^2_{MLE} = \frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{\mu})^2, \tag{118}$$

which coincides with the method-of-moments estimator (the unbiased sample variance uses denominator $M-1$ instead).

Now treat σ^2 as a random variable and assume an (improper) uniform prior on $\sigma^2 \geq 0$. Let

$$S \equiv \sum_{i=1}^{M}(x_i - \hat{\mu})^2 = M\hat{\sigma}^2_{MLE}. \tag{119}$$

The posterior is

$$\rho(\sigma^2) = Z^{-1}(\sigma^2)^{-M/2}\exp\left[-\frac{S}{2\sigma^2}\right]. \tag{120}$$

This is an inverse-gamma distribution. Writing

$$\alpha = \frac{M}{2} - 1, \qquad \beta = \frac{S}{2}, \tag{121}$$

the normalization constant is

$$Z = \int_0^\infty (\sigma^2)^{-M/2}\exp\left[-\frac{S}{2\sigma^2}\right]\, d\sigma^2 = \Gamma\left(\frac{M}{2}-1\right)\left(\frac{2}{S}\right)^{\frac{M}{2}-1}, \tag{122}$$

The common summary statistics (when they exist) are:

$$\langle\sigma^2\rangle = \frac{\beta}{\alpha - 1} = \frac{S/2}{\frac{M}{2}-2} = \frac{S}{M-4} \quad (\text{for } M > 4), \tag{123}$$

$$\text{mode}(\sigma^2) = \frac{\beta}{\alpha+1} = \frac{S/2}{\frac{M}{2}} = \frac{S}{M} = \hat{\sigma}^2_{MLE} \quad (\text{for } M \geq 2), \tag{124}$$

$$\text{Var}(\sigma^2) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} = \frac{(S/2)^2}{\left(\frac{M}{2}-2\right)^2\left(\frac{M}{2}-3\right)} = \frac{S^2}{4\left(\frac{M}{2}-2\right)^2\left(\frac{M}{2}-3\right)} \quad (\text{for } M > 6). \tag{125}$$

The likelihood (viewed as a function of μ) is a Gaussian

$$L(\mu) = \sqrt{\frac{M}{2\pi\sigma^2}}\exp\left[-\frac{M}{2\sigma^2}(\mu - \bar{x})^2\right], \qquad \bar{x} = \frac{1}{M}\sum_{i=1}^{M} x_i. \tag{126}$$

With the uniform prior the posterior is proportional to the likelihood,thus the posterior summaries are

$$\langle\mu\rangle_{\text{post}} = \bar{x}, \qquad \text{mode}(\mu) = \bar{x}, \qquad \text{Var}(\mu) = \frac{\sigma^2}{M}. \tag{127}$$

Power Law

See complete discussion in Appendix B of Newman, M. E. J. Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46, 323–351 (2005).

VIII.2. Solve the German tank problem using both MLE and method of moments estimators. Assume sampling without replacement from the integer set $\{1, 2, \ldots, M\}$. Derive estimators for M and compute their expected errors. *hint:* The likelihood reads $L(s_1, s_2, \ldots, s_n|N) = \frac{1}{\binom{N}{n}} I(s_n \leq N)$, where $I(A) = 1$ if A is true and equals zero otherwise.

Since the likelihood is discrete, we cannot maximize it by derivation. However, we can solve the problem noting that the likelihood is a decreasing function of N when $N > s_n$,

$$L(s_1, s_2, \ldots, s_n|N_1) > L(s_1, s_2, \ldots, s_n|N_2), \quad \text{if } s_n > N_2 > N_1. \tag{128}$$

Therefore, on the one hand, to maximize the likelihood, N should be the biggest possible. On the other hand, if $N < s_n$, then $L(s_1, s_2, \ldots, s_n|N < s_n) = 0$. Therefore, the maximum likelihood is obtained when N equals s_n, giving rise to the MLE

$$\hat{N}_{MLE} = s_n. \tag{129}$$

IX. STOCHASTIC PROCESSES

IX.1. Defining the probability current in the Fokker–Planck equation as

$$J(x) = A(x)\rho_t(x) - \frac{1}{2}\,\partial_x(B(x)\rho_t(x))\,, \tag{130}$$

show that the Fokker–Planck equation has the structure of a conservation law for probability: the change of probability in a region of space equals the net probability flux into that region.

The Fokker-Planck equation reads

$$\partial_t \rho_t(x) = -\partial_x J(x,t). \tag{131}$$

Integrating the left-hand side of the above equation over the interval $x \in [a, b]$, we obtain

$$\int_a^b dx\, \partial_t \rho_t(x) = \partial_t \int_a^b dx\, \rho_t(x) = \partial_t P\left(X_t \in [a, b]\right). \tag{132}$$

Performing the same operation on the right-hand side, we find

$$\int_a^b dx\ (-\partial_x J(x,t)) = J(a,t) - J(b,t). \tag{133}$$

Therefore, the time variation of the probability in the interval $[a,b]$ equals the difference between the probability current entering and exiting the interval,

$$\partial_t P\left(X_t \in [a,b]\right) = J(a,t) - J(b,t). \tag{134}$$

IX.2. Consider a random walker that, at each time step Δt, "jumps" to the right with probability p and to the left with probability q. Each jump has fixed size Δx. Derive an equation for the probability of finding the walker at position x at time t, given that it was at x_0 at time 0 (assume t is an integer multiple of Δt). Then find the associated Fokker–Planck equation and SDE in the limit $\Delta t, \Delta x \to 0$. Let X_t be the position of the walker at time t. Considering that the position of the walker at the initial time is known ($P(X_0 = x) = \delta(x - x_0)$), we want to obtain an equation for

$$p_t(x) = P(X_t = x \mid X_0 = x_0). \tag{135}$$

By marginalization,

$$p_t(x) = \sum_{x'} P(X_t = x, X_{t-\Delta t} = x' \mid X_0 = x_0). \tag{136}$$

Using the definition of conditional probabilities,

$$p_t(x) = \sum_{x'} P(X_t = x \mid X_{t-\Delta t} = x', X_0 = x_0) P(X_{t-\Delta t} = x' \mid X_0 = x_0). \tag{137}$$

Assuming that the process is Markovian,

$$P(X_t = x \mid X_{t-\Delta t} = x', X_0 = x_0) = P(X_t = x \mid X_{t-\Delta t} = x'), \tag{138}$$

and therefore

$$p_t(x) = \sum_{x'} P(X_t = x \mid X_{t-\Delta t} = x')\, p_{t-\Delta t}(x'). \tag{139}$$

There are only three possible transitions,

$$P(X_t = x \mid X_{t-\Delta t} = x') = \begin{cases} p, & \text{if } x = x' - \Delta x, \\ q, & \text{if } x = x' + \Delta x, \\ 1 - p - q, & \text{if } x = x'. \end{cases} \tag{140}$$

Therefore,

$$p_t(x) = p\, p_{t-\Delta t}(x-\Delta x) + q\, p_{t-\Delta t}(x+\Delta x) + (1-p-q)\, p_{t-\Delta t}(x). \tag{141}$$

Rewriting this expression, we obtain a discrete continuity equation, where the change of probability at position x equals the probability flux into x minus the probability flux out of x,

$$p_t(x) - p_{t-\Delta t}(x) = p\, p_{t-\Delta t}(x-\Delta x) + q\, p_{t-\Delta t}(x+\Delta x) - (p+q)\, p_{t-\Delta t}(x). \tag{142}$$

We now take the continuum limit,

$$p_{t-\Delta t}(x) = p_t(x) - \partial_t p_t(x)\, \Delta t + \mathcal{O}((\Delta t)^2), \tag{143}$$

$$p_{t-\Delta t}(x \pm \Delta x) = p_t(x) \pm \partial_x p_t(x)\, \Delta x + \frac{1}{2}\partial_x^2 p_t(x)\, (\Delta x)^2 + \mathcal{O}((\Delta x)^3, \Delta t). \tag{144}$$

Therefore,

$$\partial_t p_t(x) = -v\, \partial_x p_t(x) + \frac{D^2}{2}\partial_x^2 p_t(x). \tag{145}$$

where

$$v = \frac{(p-q)\Delta x}{\Delta t}, \qquad D^2 = \frac{(p+q)(\Delta x)^2}{\Delta t}. \tag{146}$$

IX.3. Show that

$$\rho(x) = \frac{1}{\sqrt{2\pi D^2 t}} \exp\left[-\frac{(x-x_0)^2}{2D^2 t}\right] \tag{147}$$

is the solution of the diffusion equation

$$\partial_t P_t(x) = \frac{D^2}{2}\, \partial_x^2 P_t(x). \tag{148}$$

Let us compute the time derivative of $\rho(x)$. Writing $\rho(x)$ explicitly as a function of t,

$$\rho(x) = (2\pi D^2 t)^{-1/2} \exp\left[-\frac{(x-x_0)^2}{2D^2 t}\right], \tag{149}$$

we obtain

$$\partial_t \rho(x) = \rho(x) \left[-\frac{1}{2t} + \frac{(x-x_0)^2}{2D^2 t^2}\right]. \tag{150}$$

We now compute the second spatial derivative. The first derivative is

$$\partial_x \rho(x) = -\frac{x-x_0}{D^2 t}\, \rho(x), \tag{151}$$

and the second derivative is therefore

$$\partial_x^2 \rho(x) = \left[\frac{(x - x_0)^2}{D^4 t^2} - \frac{1}{D^2 t} \right] \rho(x). \tag{152}$$

Multiplying by $D^2/2$, we find

$$\frac{D^2}{2} \partial_x^2 \rho(x) = \rho(x) \left[-\frac{1}{2t} + \frac{(x - x_0)^2}{2D^2 t^2} \right]. \tag{153}$$

Comparing this expression with $\partial_t \rho(x)$, we conclude that

$$\partial_t \rho(x) = \frac{D^2}{2} \partial_x^2 \rho(x), \tag{154}$$

which shows that $\rho(x)$ is indeed a solution of the diffusion equation.

IX.4. Using the previous result, find the solution to

$$\partial_t P_t(x) = -v \, \partial_x P_t(x) + \frac{D^2}{2} \partial_x^2 P_t(x). \tag{155}$$

Hint: use an appropriate change of variables to map this equation to the diffusion equation with known solution. What physical effect does the parameter v control?

We want to solve

$$\partial_t P_t(x) = -v \, \partial_x P_t(x) + \frac{D^2}{2} \partial_x^2 P_t(x). \tag{156}$$

We perform a change of variables that removes the drift term. Let us define

$$y = x - vt, \qquad P_t(x) = \tilde{P}_t(y). \tag{157}$$

Using the chain rule, the derivatives transform as

$$\partial_t P_t(x) = \partial_t \tilde{P}_t(y) - v \, \partial_y \tilde{P}_t(y), \tag{158}$$

$$\partial_x P_t(x) = \partial_y \tilde{P}_t(y), \qquad \partial_x^2 P_t(x) = \partial_y^2 \tilde{P}_t(y). \tag{159}$$

Substituting these expressions into the original equation, we obtain

$$\partial_t \tilde{P}_t(y) - v \, \partial_y \tilde{P}_t(y) = -v \, \partial_y \tilde{P}_t(y) + \frac{D^2}{2} \partial_y^2 \tilde{P}_t(y), \tag{160}$$

which simplifies to

$$\partial_t \tilde{P}_t(y) = \frac{D^2}{2} \partial_y^2 \tilde{P}_t(y). \tag{161}$$

This is the diffusion equation, whose solution with initial condition $\tilde{P}_0(y) = \delta(y - x_0)$ is

$$\tilde{P}_t(y) = \frac{1}{\sqrt{2\pi D^2 t}} \exp\left[-\frac{(y - x_0)^2}{2D^2 t}\right]. \tag{162}$$

Returning to the original variable x, we find

$$P_t(x) = \frac{1}{\sqrt{2\pi D^2 t}} \exp\left[-\frac{(x - x_0 - vt)^2}{2D^2 t}\right]. \tag{163}$$

The parameter v controls the drift velocity of the probability distribution: it shifts the center of the Gaussian linearly in time, describing a systematic motion superimposed on diffusion.
