Summary and problems: Probability and statistical thinking

Javier Aguilar

I. PROBABILITY SPACES: DISCRETE AND CONTINUOUS

A random variable X is defined over a sample space Ω which lists all the possible outcomes. Depending on the nature of Ω, random variables may be either discrete or continuous.

For a discrete random variable X taking values in a countable sample space $\Omega = \{x_1, x_2, \dots\}$, probabilities are assigned via a probability mass function (pmf):

$$P(X = x) = p(x), \qquad \text{with } \sum_{x\in\Omega} p(x) = 1. \tag{1}$$

For a continuous random variable X taking values in an interval or region $\Omega \subseteq \mathbb{R}$, probabilities are assigned using a probability density function (pdf):

$$P(X \in A) = \int_A \rho(x)\, dx, \qquad \text{with } \int_\Omega \rho(x)\, dx = 1. \tag{2}$$

Unlike the discrete case, the probability of X taking any particular exact value is zero:

$$P(X = x) = 0, \tag{3}$$

and only probability over intervals makes sense.

Examples:

- Discrete: outcomes of a die roll, $\Omega = \{1, 2, 3, 4, 5, 6\}$, with $P(X = i) = 1/6$.
- Continuous: measurement of human heights, that typically follow Gaussian distributions, $P(X \in [x, x + dx]) = \sqrt{\frac{1}{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

When working with multiple random variables, their combined probability is described by a joint distribution. For discrete variables X, Y:

$$P(X = x, Y = y), \tag{4}$$

while for continuous variables:

$$P((X, Y) \in A) = \iint_A \rho(x, y)\, dx\, dy. \tag{5}$$

Problems

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

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

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

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 $P(X > 0.5)$.

II. EXPECTED VALUES AND MOMENTS

Given a random variable X defined on a sample space Ω, the mean (also called expectation or first moment) is:

$$\langle X \rangle = \begin{cases} \sum_{x \in \Omega} x\, P(X = x), & \text{if } X \text{ is discrete,} \\ \int_\Omega x\, \rho(x)\, dx, & \text{if } X \text{ is continuous.} \end{cases} \tag{6}$$

The variance (also called the second central moment) is:

$$\mathrm{Var}(X) = \left\langle \big(X - \langle X \rangle\big)^2 \right\rangle = \begin{cases} \sum_{x \in \Omega} x^2\, P(X = x) - \langle X \rangle^2, & \text{if } X \text{ is discrete,} \\ \int_\Omega x^2\, \rho(x)\, dx - \langle X \rangle^2, & \text{if } X \text{ is continuous.} \end{cases} \tag{7}$$

Notation: the following symbols denote the same quantity:

$$\mathrm{Var}(X) = \left\langle \big(X - \langle X \rangle\big)^2 \right\rangle. \tag{8}$$

The mean represents the "center of mass" of the probability distribution, while the variance quantifies the typical spread or deviation of the distribution from the mean.

In general, for a transformed random variable $Y = g(X)$, the expected value is:

$$\langle Y \rangle = \begin{cases} \sum_{x \in \Omega} g(x)\, P(X = x), & \text{if } X \text{ is discrete,} \\ \int_\Omega g(x)\, \rho(x)\, dx, & \text{if } X \text{ is continuous.} \end{cases} \tag{9}$$

Problems

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

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

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

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

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

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 $x \geq x_{\min}$.

III. JOINT AND CONDITIONED PROBABILITIES, AND BAYES THEOREM

Joint probability: Given two random variables X and Y defined over sample spaces Ω_X and Ω_Y, their joint probability distribution is defined as

$$P(X = x, Y = y), \tag{10}$$

which gives the probability that $X = x$ and $Y = y$ simultaneously. If X and Y are independent, then

$$P(X = x, Y = y) = P(X = x)P(Y = y). \tag{11}$$

The joint and marginal distributions are related through marginalization,

$$P(X = x) = \sum_y P(X = x, Y = y), \quad P(Y = y) = \sum_x P(X = x, Y = y). \tag{12}$$

Conditional probability: The conditional probability of A given B is defined as

$$P(A|B) = \frac{P(A, B)}{P(B)}, \qquad P(B) \neq 0. \tag{13}$$

Bayes' theorem: How to invert condition probabilities

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{14}$$

Problems

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

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

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

$$\sum_A P(A|B) = 1. \tag{15}$$

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

IV. GALTON'S THEORY OF HUMAN HEIGHT HERITAGE

Galton measured the heights of parents and their children (denoted by h_p and h_c, respectively) and studied their deviations from the population mean, defined as $y_p = h_p - \langle h_p \rangle$ and $y_c = h_c - \langle h_c \rangle$. His central finding was the *regression to the mean*: on average, the deviation of the children is smaller than that of their parents,

$$\langle y_c \rangle = \frac{2}{3} \langle y_p \rangle.$$

Problems

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.

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

$$P(h_P \leq 72 \text{ inches}) = 0.9, \tag{16}$$

$$P(h_C > 72 \text{ inches}) = 0.118, \tag{17}$$

$$P(h_C > 72.0 \text{ inches} | h_P > 72.0 \text{ inches}) = 0.556. \tag{18}$$

show that Galton's claim is correct.

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

The goal of probability theory is to provide methods for estimating expected outcomes and statistical behavior of random processes *a priori*, without empirical data. However,

probability theory itself does not prescribe how probabilities should be assigned. In this section, we discuss probabilities of discrete random variables.

Maximum ignorance: also called the symmetric guess or maximum entropy principle. If X takes values in the sample space Ω containing n possible outcomes, and no further information is available, X is assumed to be uniformly distributed:

$$P(X = x) = \frac{1}{n}, \quad x \in \Omega. \tag{19}$$

For example, for a fair die with 6 faces, $\Omega = \{1, 2, 3, 4, 5, 6\}$, and if there is no evidence of bias, we assume $P(X = i) = \frac{1}{6}$ for all $i \in \Omega$.

Frequentist definition of probabilities: also called the classical definition. If all elementary outcomes are equally likely, then for an event A:

$$P(X \in A) = \frac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}} = \frac{n_A}{n}. \tag{20}$$

Computing n_A and n reduces to counting. For example, if we roll two dice and sum their results, and we ask for the probability that the result is 8 (i.e. $A = \{8\}$), we must count how many possible outcomes yield the total 8 and how many total outcomes exist. Exhaustive enumeration is possible, but impractical when many possibilities arise. Instead, we apply standard counting rules:

- *Product rule (with replacement):* If the outcomes arise from r independent extractions from a fixed set of m possibilities, then

$$n = m^r. \tag{21}$$

- *Permutations:* The number of possible orderings of m distinct elements is

$$n = m! = m(m-1)(m-2)\dots 2, \quad \text{with } 1! = 0! = 1. \tag{22}$$

- *Sampling without replacement (ordered):* If r distinct elements are selected in order from a set of m elements, then

$$n = m(m-1)(m-2)\dots(m-r+1) = \frac{m!}{(m-r)!}, \quad \text{with } m \geq r. \tag{23}$$

- *Sampling without replacement (unordered):* If r distinct elements are selected from m, ignoring their order, then

$$n = \binom{m}{r} = \frac{m!}{r!(m-r)!}. \tag{24}$$

- *Gombaud's problem:* Which of the following games is more likely to win?

  Game A: win if, in four throws of a single die, at least one six appears.

  Game B: win if, in 24 throws of two dice, at least one double six appears.

Problems

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?

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.

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.

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,

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i, \tag{25}$$

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

$$\lim_{N \to \infty} \hat{\mu} = \langle X \rangle. \tag{26}$$

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

$$Y = \frac{1}{N} \sum_{i=1}^{N} X^{(i)}, \tag{27}$$

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

$$\rho(y)\, dy = P(Y \in [y, y + dy]) = \sqrt{\frac{N}{2\pi\sigma^2}}\, \exp\left[-N \frac{(y-\mu)^2}{2\sigma^2} \right] dy. \tag{28}$$

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,

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

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.

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. Finally: how many measurements are required so that the empirical mean agrees with the true mean to 99% confidence?

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:* $X \sim \mathrm{Binomial}(N, p)$. 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.

Useful forms

Binomial distribution: If $X \sim \text{Binomial}(N, p)$, then

$$P(X = k) = \binom{N}{k} p^k (1-p)^{N-k}. \tag{29}$$

Poisson distribution: If $X \sim \text{Poisson}(\lambda)$, then

$$P(X = k) = \frac{\lambda^k}{k!}\, e^{-\lambda}. \tag{30}$$

Stirling's approximation: for large n,

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n, \tag{31}$$

or equivalently,

$$\ln(n!) \sim n\ln n - n + \frac{1}{2}\ln(2\pi n). \tag{32}$$

VII. THE MONTY HALL PARADOX

In the Monty Hall game, a contestant is presented with three closed doors. Behind one door there is a prize (e.g. a car), and behind the other two there is no prize. The contestant first chooses one door. Then the host, who knows where the prize is, opens one of the remaining two doors, always revealing a door without a prize. The contestant is then allowed either to keep the original choice or to switch to the other unopened door.

The counterintuitive result is that switching doors increases the probability of winning: the probability of winning by staying with the original door is 1/3, while the probability of winning by switching is 2/3.

Problems

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

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).

Inference: Parameter estimation of probability distributions

Method of moments: The method of moments estimates the parameters of a distribution by matching the empirical moments to their corresponding theoretical moments. If the distribution has k parameters $\theta_1, \ldots, \theta_k$, then one enforces

$$\frac{1}{N}\sum_{i=1}^{N} X_i^r = \langle X^r \rangle_\theta, \qquad r = 1, \ldots, k, \tag{33}$$

and solves for the θ_j.

Maximum likelihood estimation (MLE): Given data $\{x_1, \ldots, x_N\}$ drawn independently from a distribution with parameter(s) θ, the likelihood is

$$L(\theta) = \prod_{i=1}^{N} p(x_i|\theta). \tag{34}$$

The MLE estimator is

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta \log L(\theta). \tag{35}$$

Bayesian estimation: In Bayesian inference the parameters are treated as random variables with prior distribution $P(\theta)$. Given data D, one computes the posterior

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}. \tag{36}$$

Typical estimators are the posterior mean $\langle \theta|D \rangle$ and the mode of $P(\theta|D)$.

Problems

VII.1. Using both the method of moments and maximum likelihood estimation, derive the parameter estimators for the Bernoulli, 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.

VII.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.

Stochastic processes

Stochastic processes are random variables labeled by time

$$X \longrightarrow X_t, \tag{37}$$

meaning that the statistical properties of X may change over time. We will study *Markovian* processes, meaning that past information is irrelevant when conditioning on the "present,"

$$P(X_t \in A \mid X_s \in B,\, X_{s'} \in C) = P(X_t \in A \mid X_s \in B), \qquad \forall\, t \geq s > s'. \tag{38}$$

When both the process (X_t) and time are continuous, there are two main approaches to studying stochastic processes. The first is through the stochastic differential equation

$$\dot{X}_t = A(X_t) + B(X_t)\,\xi(t), \tag{39}$$

where A and B are the drift and diffusion functions, representing deterministic and random contributions to the dynamics of X_t. The term $\xi(t)$ is called "white noise," a random process defined through the statistics

$$\int_t^{t+\Delta t} \xi(s)\,ds \sim G(0, \Delta t), \qquad \langle \xi(s)\,\xi(s') \rangle = \delta(s - s'), \tag{40}$$

where $G(0, \Delta t)$ denotes a Gaussian random variable with mean 0 and variance Δt.

Alternatively, one may study the dynamics of the probability distribution of X_t, called the propagator,

$$\rho(x) = \lim_{dx \to 0} \frac{1}{dx} P\big(X_t \in [x, x + dx] \mid X_0 = x_0\big), \tag{41}$$

which satisfies the Fokker–Planck equation

$$\partial_t \rho_t(x) = -\partial_x (A(x)\rho_t(x)) + \frac{1}{2}\partial_x^2 (B(x)\rho_t(x))\,. \tag{42}$$

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

$$J(x) = A(x)\rho_t(x) - \frac{1}{2}\,\partial_x (B(x)\rho_t(x))\,, \tag{43}$$

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.

VII.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 $\Delta t, \Delta x \to 0$.

VII.3. Show that

$$\rho(x) = \frac{1}{\sqrt{2\pi D^2 t}} \exp\left[-\frac{(x - x_0)^2}{2D^2 t}\right] \tag{44}$$

is the solution of the diffusion equation

$$\partial_t P_t(x) = \frac{D^2}{2} \partial_x^2 P_t(x). \tag{45}$$

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

$$\partial_t P_t(x) = -v \, \partial_x P_t(x) + \frac{D^2}{2} \partial_x^2 P_t(x). \tag{46}$$

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?
