Summary and problems: Probability and statistical thinking

Javier Aguilar

I. EXPECTED VALUES AND MOMENTS

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

$$\langle X\rangle = \begin{cases} \sum_{x\in\Omega} x\,P(X=x), & \text{if } X \text{ is discrete,} \\ \int_\Omega x\,\rho(x)\,dx, & \text{if } X \text{ is continuous.} \end{cases} \tag{1}$$

Notation: the following symbols denote the same quantity:

$$\mu = E[X] = \langle X\rangle. \tag{2}$$

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

$$\mathrm{Var}(X) = \left\langle \big(X - \langle X\rangle\big)^2 \right\rangle = \begin{cases} \sum_{x\in\Omega} x^2\,P(X=x) - \langle X\rangle^2, & \text{if } X \text{ is discrete,} \\ \int_\Omega x^2\,\rho(x)\,dx - \langle X\rangle^2, & \text{if } X \text{ is continuous.} \end{cases} \tag{3}$$

Notation: the following symbols denote the same quantity:

$$\sigma^2 = \mathrm{Var}(X) = \left\langle \big(X - \langle X\rangle\big)^2 \right\rangle. \tag{4}$$

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:

$$\langle Y\rangle = \begin{cases} \sum_{x\in\Omega} g(x)\,P(X=x), & \text{if } X \text{ is discrete,} \\ \int_\Omega g(x)\,\rho(x)\,dx, & \text{if } X \text{ is continuous.} \end{cases} \tag{5}$$

Problems

I.1. Show that $var(X) = \langle X^2\rangle - \mu^2$.

I.2. Compute $\langle\langle Y\rangle\rangle$. Also, what is the expected value of a constant?

I.3. Compute the mean and variance of a

II. GALTON'S THEORY OF HUMAN HEIGHT HERITAGE

Galton measured the height of parents and their children (h_p and h_c). He measured the mean of both heights, called μ_p and μ_c. Then he studied the variation with respect to the mean: $y_p = h_p - \mu_p$ and $y_c = h_c - \mu_c$. His main result is the regression to the mean: the average variation of children is smaller than the one of the parents:

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

Problems

II.1. Galton's theory of heritage can be expressed using the figure of the "generant": with probability p, children inherit the height deviation of one of their parents; with probability $1 - p$, children inherit the height of the "generant", whose height equals the population mean. Given Galton's regression ratio of 2/3, what is the value of p?

II.2. Prove Galton's claim: tall parents tend to have tall children, but tall children do not necessarily have tall parents. *Hint:* Let T_p denote the event that parents are tall (positive deviation from the mean) and T_c the event that a child is tall. Use Bayes' theorem to compare $P(T_c \mid T_p)$ and $P(T_p \mid T_c)$.

III. 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:

$$P(X = x) = \frac{1}{n}, \quad x \in \Omega. \tag{6}$$

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:

$$P(X \in A) = \frac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}} = \frac{n_A}{n}. \tag{7}$$

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

$$n = m^r. \tag{8}$$

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

$$n = m! = m(m-1)(m-2)\dots 2, \quad \text{with } 1! = 0! = 1. \tag{9}$$

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

$$n = m(m-1)(m-2)\dots(m-r+1) = \frac{m!}{(m-r)!}, \quad \text{with } m \geq r. \tag{10}$$

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

$$n = \binom{m}{r} = \frac{m!}{r!(m-r)!}. \tag{11}$$

- *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

III.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?

III.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.

III.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.

IV. 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,

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{12}$$

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

$$\lim_{N\to\infty} \hat{\mu} = \langle X \rangle. \tag{13}$$

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

$$Y = \frac{1}{N}\sum_{i=1}^{N} X^{(i)}, \tag{14}$$

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

$$\rho(y)\,dy = P(Y \in [y, y+dy]) = \sqrt{\frac{N}{2\pi\sigma^2}}\,\exp\left[-N\frac{(y-\mu)^2}{2\sigma^2}\right] dy. \tag{15}$$

Problems

IV.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,

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

IV.2. *** Show that the central limit theorem applies to the empirical estimator of the mean. Show explicitly that the central limit theorem implies the law of large numbers.

IV.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?

IV.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.

Useful forms

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

$$P(X = k) = \binom{N}{k} p^k (1-p)^{N-k}. \tag{16}$$

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

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{17}$$

Stirling's approximation: for large n,

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n, \tag{18}$$

or equivalently,

$$\ln(n!) \sim n \ln n - n + \frac{1}{2} \ln(2\pi n). \tag{19}$$

V. 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

$$P(X = x, Y = y), \tag{20}$$

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

$$P(X = x, Y = y) = P(X = x)P(Y = y). \tag{21}$$

The joint and marginal distributions are related through marginalization,

$$P(X = x) = \sum_y P(X = x, Y = y), \quad P(Y = y) = \sum_x P(X = x, Y = y). \tag{22}$$

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

$$P(A|B) = \frac{P(A, B)}{P(B)}, \qquad P(B) \neq 0. \tag{23}$$

Bayes' theorem: How to invert condition probabilities

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{24}$$

Problems

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

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

V.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 $P(B|A)$.

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

$$\sum_A P(A|B) = 1. \tag{25}$$

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

VI. THE MONTY HALL PARADOX

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

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

VI.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

$$\frac{1}{N}\sum_{i=1}^{N} X_i^r = \langle X^r \rangle_\theta, \qquad r = 1, \ldots, k, \tag{26}$$

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

$$L(\theta) = \prod_{i=1}^{N} p(x_i|\theta). \tag{27}$$

The MLE estimator is

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta \log L(\theta). \tag{28}$$

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

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}. \tag{29}$$

Typical estimators are the posterior mean $\mathbb{E}[\theta|D]$ and the MAP estimate $\arg\max_\theta P(\theta|D)$.

Problems

VI.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.

VI.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.