

1.S: Summary

References

- C. Gardiner, *Stochastic Methods* (4th edition, Springer-Verlag, 2010) Very clear and complete text on stochastic methods with many applications.
- J. M. Bernardo and A. F. M. Smith, *Bayesian Theory* (Wiley, 2000) A thorough textbook on Bayesian methods.
- D. Williams, *Weighing the Odds: A Course in Probability and Statistics* (Cambridge, 2001) A good overall statistics textbook, according to a mathematician colleague.
- E. T. Jaynes, *Probability Theory* (Cambridge, 2007) An extensive, descriptive, and highly opinionated presentation, with a strongly Bayesian approach.
- A. N. Kolmogorov, *Foundations of the Theory of Probability* (Chelsea, 1956) The *Urtext* of mathematical probability theory.

Summary

• **Discrete distributions:** Let n label the distinct possible outcomes of a discrete random process, and let p_n be the probability for outcome n . Let A be a quantity which takes values which depend on n , with A_n being the value of A under the outcome n . Then the expected value of A is $\langle A \rangle = \sum_n p_n A_n$, where the sum is over all possible allowed values of n . We must have that the distribution is normalized, $\langle 1 \rangle = \sum_n p_n = 1$.

• **Continuous distributions:** When the random variable φ takes a continuum of values, we define the *probability density* $P(\varphi)$ to be such that $P(\varphi) d\mu$ is the probability for the outcome to lie within a differential volume $d\mu$ of φ , where $d\mu = W(\varphi) \prod_{i=1}^n d\varphi_i$, where φ is an n -component vector in the configuration space Ω , and where the function $W(\varphi)$ accounts for the possibility of different configuration space measures. Then if $A(\varphi)$ is any function on Ω , the expected value of A is $\langle A \rangle = \int_{\Omega} d\mu P(\varphi) A(\varphi)$.

• **Central limit theorem:** If $\{x_1, \dots, x_N\}$ are each independently distributed according to $P(x)$, then the distribution of the sum $X = \sum_{i=1}^N x_i$ is

$$\mathcal{P}_N(X) = \int_{-\infty}^{\infty} dx_1 \cdots \int_{-\infty}^{\infty} dx_N P(x_1) \cdots P(x_N) \delta\left(X - \sum_{i=1}^N x_i\right) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} [\hat{P}(k)]^N e^{ikX}, \quad (1.S.1)$$

where $\hat{P}(k) = \int dx P(x) e^{-ikx}$ is the Fourier transform of $P(x)$. Assuming that the lowest moments of $P(x)$ exist, $\ln[\hat{P}(k)] = -i\mu k - \frac{1}{2}\sigma^2 k^2 + \mathcal{O}(k^3)$, where $\mu = \langle x \rangle$ and $\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2$ are the mean and standard deviation. Then for $N \rightarrow \infty$,

$$P_N(X) = (2\pi N\sigma^2)^{-1/2} e^{-(X - N\mu)^2 / 2N\sigma^2}, \quad (1.S.2)$$

which is a Gaussian with mean $\langle X \rangle = N\mu$ and standard deviation $\sqrt{\langle X^2 \rangle - \langle X \rangle^2} = \sqrt{N} \sigma$. Thus, X is distributed as a Gaussian, even if $P(x)$ is not a Gaussian itself.

• **Entropy:** The entropy of a statistical distribution is $\{p_n\}$ is $S = -\sum_n p_n \ln p_n$. (Sometimes the base 2 logarithm is used, in which case the entropy is measured in *bits*.) This has the interpretation of the *information content* per element of a random sequence.

• **Distributions from maximum entropy:** Given a distribution $\{p_n\}$ subject to $(K+1)$ constraints of the form $\mathcal{X}^a = \sum_n X_n^a p_n$ with $a \in \{0, \dots, K\}$, where $\mathcal{X}^0 = X_n^0 = 1$ (normalization), the distribution consistent with these constraints which maximizes the entropy function is obtained by extremizing the multivariable function

$$S^*(\{p_n\}, \{\lambda_a\}) = -\sum_n p_n \ln p_n - \sum_{a=0}^K \lambda_a \left(\sum_n X_n^a p_n - \mathcal{X}^a \right), \quad (1.S.3)$$

with respect to the probabilities $\{p_n\}$ and the Lagrange multipliers $\{\lambda_a\}$. This results in a Gibbs distribution,

$$p_n = \frac{1}{Z} \exp \left\{ -\sum_{a=1}^K \lambda_a X_n^a \right\}, \quad (1.S.4)$$

where $Z = e^{1+\lambda_0}$ is determined by normalization, $\sum_n p_n = 1$ (the $a = 0$ constraint) and the K remaining multipliers determined by the K additional constraints.

- *Multidimensional Gaussian integral:*

$$\int_{-\infty}^{\infty} dx_1 \cdots \int_{-\infty}^{\infty} dx_n \exp\left(-\frac{1}{2} x_i A_{ij} x_j + b_i x_i\right) = \left(\frac{(2\pi)^n}{\det A}\right)^{1/2} \exp\left(\frac{1}{2} b_i A_{ij}^{-1} b_j\right). \quad (1.S.5)$$

- *Bayes' theorem:* Let the conditional probability for B given A be $P(B|A)$. Then Bayes' theorem says $P(A|B) = P(A) \cdot P(B|A) / P(B)$. If the 'event space' is partitioned as $\{A_i\}$, then we have the extended form,

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_j P(B|A_j) \cdot P(A_j)}. \quad (1.S.6)$$

When the event space is a 'binary partition' $\{A, \neg A\}$, as is often the case in fields like epidemiology (test positive or test negative), we have

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)}. \quad (1.S.7)$$

Note that $P(A|B) + P(\neg A|B) = 1$ (which follows from $\neg \neg A = A$).

- *Updating Bayesian priors:* Given data in the form of observed values $\mathbf{x} = \{x_1, \dots, x_N\} \in \mathcal{X}$ and a hypothesis in the form of parameters $\theta = \{\theta_1, \dots, \theta_K\} \in \Theta$, we write the conditional probability (density) for observing \mathbf{x} given θ as $f(\mathbf{x}|\theta)$. Bayes' theorem says that the corresponding distribution $\pi(\theta|\mathbf{x})$ for θ conditioned on \mathbf{x} is

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta) \pi(\theta)}{\int_{\Theta} d\theta' f(\mathbf{x}|\theta') \pi(\theta')}, \quad (1.S.8)$$

We call $\pi(\theta)$ the *prior* for θ , $f(\mathbf{x}|\theta)$ the *likelihood* of \mathbf{x} given θ , and $\pi(\theta|\mathbf{x})$ the *posterior* for θ given \mathbf{x} . We can use the posterior to find the distribution of new data points \mathbf{y} , called the *posterior predictive distribution*, $f(\mathbf{y}|\mathbf{x}) = \int_{\Theta} d\theta f(\mathbf{y}|\theta) \pi(\theta|\mathbf{x})$. This is the update of the *prior predictive distribution*, $f(\mathbf{x}) = \int_{\Theta} d\theta f(\mathbf{x}|\theta) \pi(\theta)$. As an example, consider coin flipping with $f(\mathbf{x}|\theta) = \theta^X (1-\theta)^{N-X}$, where N is the number of flips, and $X = \sum_{j=1}^N x_j$ with x_j a discrete variable which is 0 for tails and 1 for heads. The parameter $\theta \in [0, 1]$ is the probability to flip heads. We choose a prior $\pi(\theta) = \theta^{\alpha-1} (1-\theta)^{\beta-1} / B(\alpha, \beta)$ where $B(\alpha, \beta) = \Gamma(\alpha) \Gamma(\beta) / \Gamma(\alpha + \beta)$ is the Beta distribution. This results in a normalized prior $\int_0^1 d\theta \pi(\theta) = 1$. The posterior distribution for θ is then

$$\pi(\theta|x_1, \dots, x_N) = \frac{f(x_1, \dots, x_N|\theta) \pi(\theta)}{\int_0^1 d\theta' f(x_1, \dots, x_N|\theta') \pi(\theta')} = \frac{\theta^{X+\alpha-1} (1-\theta)^{N-X+\beta-1}}{B(X+\alpha, N-X+\beta)}. \quad (1.S.9)$$

The prior predictive is $f(\mathbf{x}) = \int_0^1 d\theta f(\mathbf{x}|\theta) \pi(\theta) = B(X+\alpha, N-X+\beta) / B(\alpha, \beta)$, and the posterior predictive for the total number of heads Y in M flips is

$$f(\mathbf{y}|\mathbf{x}) = \int_0^1 d\theta f(\mathbf{y}|\theta) \pi(\theta|\mathbf{x}) = \frac{B(X+Y+\alpha, N-X+M-Y+\beta)}{B(X+\alpha, N-X+\beta)}. \quad (1.S.10)$$

This page titled [1.S: Summary](#) is shared under a [CC BY-NC-SA](#) license and was authored, remixed, and/or curated by [Daniel Arovas](#).