Due date: 27 January 2023 (2 PM) The first three and half pages of this document provides you with notes on distribution functions for your perusal (Source: "Molecular Driving Forces" by Ken Dill). Homework questions start after that.

Distribution Functions Have Average Values and Standard Deviations

Averages

A probability distribution function contains all the information that can be known about a probabilistic system. A full distribution function, however, is rarely accessible from experiments. Generally, experiments can measure only certain moments of the distribution. The nth moment of a probability distribution function $p(x)$ is

$$<x^n> = \int_a^b x^n p(x)dx = \frac{\int_a^b x^n g(x)dx}{\int_a^b g(x)dx} \tag{1}$$

where the second expression is appropriate for a non-normalized distribution function $g(x)$. Angle brackets <> are used to indicate the moments, also called the expectation values or averages, of a distribution function. For a probability distribution the zeroth moment always equals one, because the sum of the probabilities equals one. The first moment of a distribution function ($n = 1$) in Eq.1 is called the mean, average, or expected value. For discrete functions,

$$<i> = \sum_{i=1}^{t} ip(i), \tag{2}$$

and for continuous functions,

$$<x> = \int_a^b xp(x)dx \tag{3}$$

For distributions over t discrete values, the mean of a function $f(i)$ is

$$<f(i)> = \sum_{i=1}^{t} f(i)p(i) \tag{4}$$

For distributions over continuous values, the mean of a function $f(x)$ is

$$<f(x)> = \int_a^b f(x)p(x)dx = \frac{\int_a^b f(x)g(x)dx}{\int_a^b g(x)dx} \tag{5}$$

Here are two useful and general properties of averages, derived from the definition given in Eq. 5:

$$\begin{aligned} <af(x)> = \int af(x)p(x)dx &= a\int f(x)p(x)dx \\ &= a<f(x)>, \end{aligned} \tag{6}$$

where a is a constant.

$$\begin{aligned} <f(x)+g(x)> &= \int [f(x)+g(x)]p(x)dx \\ &= \int f(x)p(x)dx + \int g(x)p(x)dx \\ &= <f(x)> + <g(x)> \end{aligned} \tag{7}$$

Variance

The variance σ^2 is a measure of the width of a distribution. A broad, flat distribution has a large variance, while a narrow, peaked distribution has a small variance. The variance σ^2 is defined as the average square deviation from the mean,

$$\sigma^2 = <(x-a)^2> = <x^2 - 2ax + a^2>, \tag{8}$$

where $a = <x>$ is the mean value, or first moment. We use a instead of $<x>$ as a reminder here that this quantity is just a constant, not a variable. Using Eq. 7, 8 becomes

$$\sigma^2 = <x^2> - <2ax> + <a^2>. \tag{9}$$

Using Eq. 6,

$$\sigma^2 = <x^2> - 2a<x> + a^2 = <x^2> - <x>^2. \tag{10}$$

Examples of Distributions

Bernoulli

$$\begin{aligned} g(n) &= p^n(1-p)^{N-n} \\ n &= 0, 1, 2, ..., N. \end{aligned} \tag{11}$$

The Bernoulli distribution describes independent trials with two possible outcomes. $g(n)$ is a distribution function, not a probability, because it is not normalized to sum to one (Fig. 1 (a)).

Poisson

$$\begin{aligned} p(n) &= \frac{a^n e^{-a}}{n!} \\ n &= 0, 1, 2, ..., N. \end{aligned} \tag{12}$$

The Poisson distribution approximates the binomial distribution when the number of trials is large and the probability of each one is small (Fig. 1 (b)). It is useful for describing radioactive decay, the number of vacancies in the Supreme Court each year, the numbers of dye molecules taken up by small particles, or the sizes of colloidal particles.

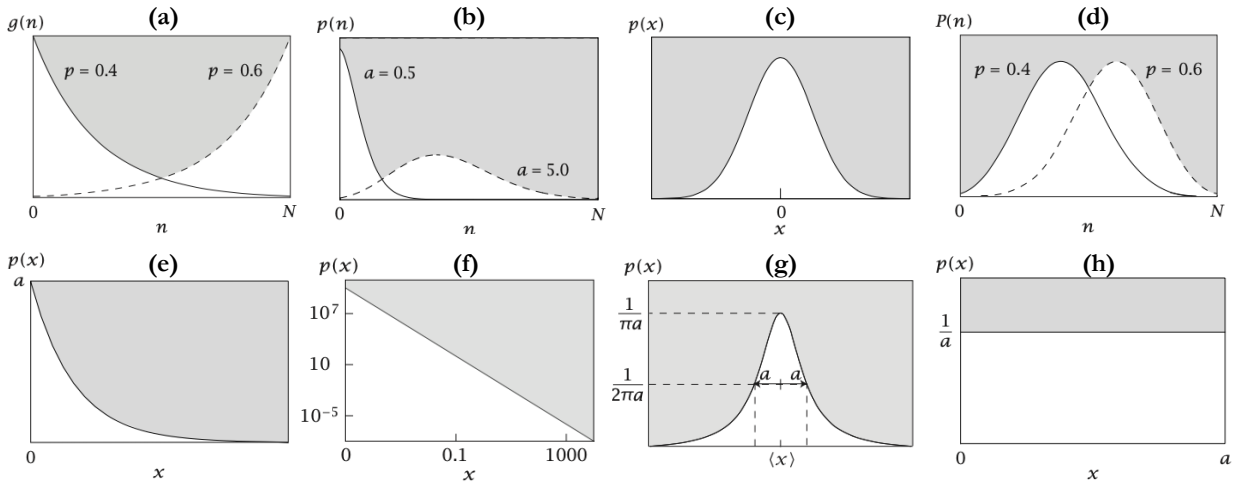

Figure 1: Probability distribution plots: (a) Bernoulli, (b) Poisson, (c) Gaussian, (d) Binomial, (e) Exponential(Boltzmann), (f) Power law, (g) Lorentzian, (h) Flat.

Gaussian

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$$
$$-\infty \leq x \leq \infty. \tag{13}$$

The Gaussian distribution is derived from the binomial distribution for large N (Fig. 1 (c)). It is important for statistics, error analysis, diffusion, conformations of polymer chains, and the Maxwell–Boltzmann distribution law of gas velocities.

Binomial

$$P(n) = p^n(1-p)^{N-n}\left(\frac{N!}{n!(N-n)!}\right),$$
$$n = 0, 1, 2, ..., N. \tag{14}$$

Exponential (Boltzmann)

$$p(x) = ae^{-ax},$$
$$0 \leq x \leq \infty. \tag{15}$$

The exponential, or Boltzmann distribution, is central to statistical thermodynamics (Fig.1 (e)).

Power law

$$p(x) = \frac{1}{x^q}, \tag{16}$$

where q is a constant called the power law exponent (Fig. 1 (f)). Power laws describe the frequencies of earthquakes, the numbers of links to World Wide Web sites, the distribution of incomes ('the rich get richer'), and the noise spectrum in some electronic devices.

Lorentzian

$$p(x) = \frac{1}{\pi}\frac{a}{(x - <x>)^2 + a^2}, \quad -\infty \leq x \leq \infty. \tag{17}$$

$2a$ is the width of the Lorentzian curve at the level of half the maximum probability (Fig. 1 (g)). Lorentzian distributions are useful in spectroscopy.

Flat

$$p(x) = \frac{1}{a}, \tag{18}$$

where a is a constant independent of x (Fig. 1 (h)).

Homework Problems

Question 1: The Maxwell–Boltzmann probability distribution function

According to the kinetic theory of gases, the energies of molecules moving along the x direction are given by $\epsilon_x = (1/2)mv_x^2$, where m is mass and v_x is the velocity in the x direction. The distribution of particles over velocities is given by the Boltzmann law, $p(v_x) = e^{-mv_x^2/2kT}$. This is the Maxwell–Boltzmann distribution (velocities may range from $-\infty$ to $+\infty$).
a) Write the probability distribution $p(v_x)$, so that the Maxwell–Boltzmann distribution is correctly normalized.
b) Compute the average energy $<\frac{1}{2}mv_x^2>$.
c) What is the average velocity $<v_x>$?
d) What is the average momentum $<mv_x>$?

Question 2: Predicting the rate of mutation based on the Poisson probability distribution function.

The evolutionary process of amino acid substitution in proteins is sometimes described by the Poisson probability distribution function. The probability $p_s(t)$ that exactly s substitutions at a given amino acid position occur over an evolutionary time t is

$$p_s(t) = \frac{e^{-\lambda t}(\lambda t)^s}{s!} \tag{19}$$

where λ is the rate of amino acid substitution per site per unit time. Fibrinopeptides evolve rapidly: $\lambda_F = 9.0$ substitutions per site per 10^9 years. Lysozyme is intermediate: $\lambda_L \approx 1.0$. Histones evolve slowly: $\lambda_H = 0.010$ substitutions per site per 10^9 years.

a) What is the probability that a fibrinopeptide has no mutations at a given site in $t = 1$ billion years?
b) What is the probability that lysozyme has three mutations per site in 100 million years?
c) We want to determine the expected number of mutations $<s>$ that will occur in time t. We will do this in two steps. First, using the fact that probabilities must sum to one, write $\alpha = \sum_{s=0}^{\infty}(\lambda t)^s/s!$ in a simpler form.
d) Now write an expression for $<s>$. Note that

$$\sum_{s=0}^{\infty}\frac{s(\lambda t)^s}{s!} = (\lambda t)\sum_{s=1}^{\infty}\frac{(\lambda t)^{s-1}}{(s-1)!} = \lambda t\alpha. \tag{20}$$

e) Using your answer to part (d), determine the ratio of the expected number of mutations in a fibrinopeptide to the expected number of mutations in histone protein, $<s>_{fib}/<s>_{his}$.

Question 3: Flat distribution.

Given a flat distribution, from $x = -a$ to $x = a$, with probability distribution $p(x) = 1/(2a)$:
a) Compute $<x>$.
b) Compute $<x^2>$.
c) Compute $<x^3>$.
d) Compute $<x^4>$.

Question 4: Family probabilities.

Given that there are three children in a family, what is the probability that:
a) two are boys and one is a girl?
b) all three are girls?

Question 5: Ion-channel events.

A biological membrane contains N ion-channel proteins. The fraction of time that any one protein is open to allow ions to flow through is q. Express the probability $P(m, N)$ that m of the channels will be open at any given time.