Published in final edited form as:

Sci Signal. 2011 September 27; 4(192): tr14. doi:10.1126/scisignal.2001980.

Probabilistic Reasoning in Data Analysis

Lawrence Sirovich*

Department of Pharmacology and Systems Therapeutics and Systems Biology Center New York,
Mount Sinai School of Medicine, New York, NY 10029, USA

Abstract

This Teaching Resource provides lecture notes, slides, and a student assignment for a lecture on probabilistic reasoning in the analysis of biological data. General probabilistic frameworks are introduced, and a number of standard probability distributions are described using simple intuitive ideas. Particular attention is focused on random arrivals that are independent of prior history (Markovian events), with an emphasis on waiting times, Poisson processes, and Poisson probability distributions. The use of these various probability distributions is applied to biomedical problems, including several classic experimental studies.

Keywords

Mathematical modeling; probability distributions; random events; stochastic phenomena; visual system; neuroscience; neuromuscular junction

Lecture Notes

This lecture introduces probabilistic reasoning and provides examples that illustrate the importance of probability in biology (1). An effort has been made to minimize mathematical details; however, it is assumed that the student is familiar with calculus.

Variability in preparations and instrument noise, features common to all biological experiments, introduce a probabilistic element into the data. In addition, probability may be inherent in the underlying biological process, so that a meaningful interpretation of the results requires considering their probabilistic nature. For instance, the end-plate potential that can be recorded at the neuromuscular junction depends on the probabilistic release of individual quanta of neurotransmitter (2, 3, 4). To model an experiment and the data that arise from experimentation, it is desirable to be able to characterize inherent trends and fluctuations in the presence of extrinsic noise. This lecture provides examples that illustrate how appropriate probabilistic reasoning can provide fundamental insight into biological phenomena, such as the size and frequency of cellular signaling events.

The lecture begins by introducing several intuitive examples of probabilistic processes, such as the probability of obtaining a particular number of heads (p_h) when flipping a fair coin a given number of times (Slides 2 and 3). After an explanation of the mathematical basis of

Copyright 2008 by the American Association for the Advancement of Science; all rights reserved.

*Corresponding author. lsirovich@mail.rockefeller.edu.

Supplementary Materials

<http://stke.sciencemag.org/cgi/content/full/sigtrans;4/192/tr14/DC1>

Slides. Probabilistic Reasoning in Data Analysis

Data file to be used for analysis in problem set.

Answer key. MATLAB code providing one solution to the problem set and accompanying PDF are available upon request.

coin flipping probability and its generalization to the example of drawing a marble of a given color out of an urn, formulas describing the binomial distribution, the gamma distribution, the Gaussian distribution, and the Poisson distribution are described (Slides 2 to 5). Directly relevant to the classic studies discussed later, the lecture demonstrates how the binomial distribution describes the probability of observing a particular number of events in a certain number of trials (Slide 2), and how the Poisson distribution can be considered a special case of the binomial distribution for rare events (Slide 5). Additionally, some examples from dice games are presented to illustrate how useful calculations can be performed with the formulas for probability distributions (Slide 6).

The next topic concerns events that occur with randomly distributed arrival times (Slide 7). This phenomenon, which is described by the Poisson distribution, is characteristic of many biologically relevant processes, including the activation of photoreceptors in the retina by photons (5). The probability of waiting times for such a Poisson process is derived from first principles (Slide 8). The lecture shows that in a Poisson process, the variance is equal to the mean (Slide 9), a property that has proven important for identifying biological processes that are governed by Poisson statistics.

A focus of the lecture is the classic experiments on the visual system of Hecht, Shlaer, and Pirenne (5) and a subsequent study by Mueller (6) (Slides 10 to 12), which together established the quantal threshold of vision under low-light conditions. A fundamental understanding of probability was critical to the success of these studies. These authors experimentally measured the probability that an observer detected a brief flash of light as a function of the flash intensity. The key insight of this study, however, occurred when the authors fit the experimental data to cumulative probability density functions on the basis of the assumption that photon absorption by the retina obeys Poisson distribution. Through these curve fits, this study determined the threshold for human detection of light. Other examples include the studies of bacterial mutations by Luria and Delbrück (7) (Slide 13) and of the quantal nature of neurotransmitter release (2, 3, 4) (Slides 13 to 15), where statistical analyses of fluctuations in the amplitude of the postsynaptic response to nerve stimulation revealed that neurotransmitter was released in “packets.”

The lecture concludes with data from the study of Hecht, Shlaer, and Pirenne (5) (Slide 12), which is used by the students in the Problem Set.

Problem Set

Introductory Details

Original data obtained by Hecht, Shlaer, and Pirenne (5) are shown in Slide 12. The file “hecht_fig7_data.txt” contains these values in tab-delimited text format. These numbers can be read into MATLAB using a command, such as the following: `data = dlmread('hecht_Fig. 7_data.txt')`. The goal of this assignment is to reproduce the analysis performed by the authors to estimate the number of photons that must be detected by the retina for the observer to perceive a flash of light.

The authors fit their data to the following formula (also shown in Slides 10 and 11):

$$C_n(a) = 1 - \sum_{k=0}^{n-1} P_k(a)$$

where

$$P_k(a) = \frac{a^k e^{-a}}{k!}$$

Student Assignment

1. Read into MATLAB the data file “hecht_Fig. 7_data.txt.” As with the Table shown in Slide 12, odd columns correspond to stimulus intensities (in units of quanta) and even columns correspond to the percentage of stimuli correctly detected (represented as fractions rather than percentages). More specifically, the file contains the 3rd and 4th, 5th and 6th, and 9th and 10th columns in the Table, which correspond to the results shown in Fig. 7 (Slide 12). Step 1 is to reproduce the plots shown in Fig. 7 (Slide 12).

2. Generate a vector a ranging from 1 to 400. Compute $C_n(a)$ according to the formula above for all values of n between 1 and 10. Plot $\log_{10}(a)$ versus $C_n(a)$, with all 10 plots on the same graph.

Two things should be obvious from comparing the experimental data with the curves generated from the equations. One is that for $n > 4$, the shape of $C_n(a)$ begins to closely resemble the shape of the experimental plots. The second is that the plots of $C_n(a)$ are considerably left-shifted compared with the experimental results.

The reason for the latter is the different units of the experimental data and the theoretical curves. The columns “No. of quanta” in figure 7 of the paper (Slide 12) represent the estimated number of photons incident at the cornea of each subject. In contrast, the x axis of the theoretical plots (that is, the vector a) represents photons detected by the retina. It is certainly reasonable to believe that only a minority of photons at the cornea actually reach, and are detected by, cells in the retina.

3. The next step in the analysis is to determine an appropriate factor to account for the percentage of photons that reach the cornea but are not detected by cells in the retina. Implementing this factor correctly will cause the plots of $C_n(a)$ to shift to the right.

Choose a scale factor. Plot the experimental data from one of the subjects. On the same graph, plot $C_n(a)$ for several values of n . Generate new plots for several different scale factors. Approximately what scale factor is required so make the experimental results approximately match the theoretical results?

Hint: Do not scale the variable that gets plotted on the x axis of the figures. These values should be the same as those listed in the Table (that is a minimum of 37.1 and a maximum of 421.7 for subject 1). Instead, scale the variable that is used to compute $C_n(a)$.

4. The final challenge is to determine the appropriate value of n for each experimental subject. Here use only the scale factor determined to be most appropriate in assignment 3.

Note that for each experiment there are six data points D_j , where $j = 1, 2, \dots, 6$. The goal is to minimize the “distance” between the data and the computed $C_n(a)$. In other words, we wish to find the value of n for which

$$Error = \sum_{j=1}^6 [D_j - C_n(a)]^2$$

is a minimum.

To complete this step, you will have to compute a different version of $C_n(a)$, this version you create is not defined at 400 values of a , but only at the same 6 values as the experimental data.

Once you have computed this, calculate the distance between D_j and $C_n(a)$ for all values of n . Plot the distance versus n to determine which value minimizes the distance.

Educational Details

Learning Resource Type: Lecture, assignment, PowerPoint

Context: Graduate

Intended Users: Teacher, learner

Intended Educational Use: Learn, plan, teach

Discipline: Biochemistry, bioengineering, biophysics, cell biology, education, physiology, theoretical biology

Technical Details

Format: PowerPoint (.ppt)

Size: 939 KB

Requirements: Microsoft PowerPoint

Format: MATLAB (.m)

Requirements: Mathworks MATLAB

Format: PDF (.pdf)

Requirements: Adobe PDF Reader

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Papoulis, A. Probability, Random Variables, and Stochastic Processes. Vol. 4. New York: McGraw-Hill; 2002.
2. Del Castillo J, Katz B. Quantal components of the end-plate potential. J. Physiol. 1954; 124:560–573. [PubMed: 13175199]
3. Katz, B. Nerve, Muscle, and Synapse. New York: McGraw-Hill; 1965.
4. Boyd IA, Martin AR. The end-plate potential in mammalian muscle. J. Physiol. 1956; 132:74–91. [PubMed: 13320373]
5. Hecht S, Schlaer S, Pirenne MH. Energy, quanta, and vision. J. Gen. Physiol. 1942; 25:819–840. [PubMed: 19873316]
6. Mueller CG. Frequency of seeing functions for intensity discrimination of various levels of adapting intensity. J. Gen. Physiol. 1951; 34:463–474. [PubMed: 14824512]
7. Luria SE, Delbrück M. Mutations of bacteria from virus sensitivity to virus resistance. Genetics. 1943; 28:491–511. [PubMed: 17247100]