Journal of Theoretical Physics

Founded and Edited by M. Apostol **14** (1996)

ISSN 1453-4428

On the empirical foundation of probability

M. Apostol
Department of Theoretical Physics,
Institute of Atomic Physics,
Magurele-Bucharest MG-6,
POBox MG-35, Romania
e-mail: apoma@theor1.ifa.ro

Abstract

It is shown that by testing an ensemble of n objects the probability can be determined with an error $\sim 1/\sqrt{n}$.

Suppose that we have N objects, out of which q have a certain feature that occurs with the probability p and the remaining $N-q$ have not that feature. Testing the whole ensemble of N objects the probability p will be given by the binomial distribution

$$f(p) = C_N^q \cdot p^q (1-p)^{N-q} \quad . \tag{1}$$

This function is positive and less than unity, because, for example,

$$\sum_{q=0}^{N} f(p) = 1 \quad . \tag{2}$$

Its first derivative

$$f'(p) = C_N^q \cdot p^{q-1}(1-p)^{N-q-1}(q-Np) \tag{3}$$

vanishes at

$$p_N = q/N \tag{4}$$

where its second derivative

$$f''(p) = C_N^q \cdot p^{q-2}(1-p)^{N-q-2}\left[N(N-1)p^2 - 2N(N-1)p_N p + Np_N(Np_N - 1)\right] \tag{5}$$

is negative,

$$f''(p_N) = -N \cdot C_N^q \cdot p_N^{q-1}(1-p_N)^{N-q-1} \quad . \tag{6}$$

Using $n! \sim n^n$ for large n, it is easy to show that $f(p_N)$ goes to unity for q and N large enough, and

$$f''(p_N) \cong -\frac{N}{p_N(1-p_N)} \rightarrow -\infty \quad . \tag{7}$$

In addition, making use of (5), the second derivative vanishes at $p \cong p_N \pm \sqrt{p_N(1-p_N)/N}$. For N large enough the distribution $f(p)$ is sharply peaked at $p = p_N$, where it approaches unity. One can say therefore that for large N the probability p is given by the empirical probability p_N with an error

$$\delta p \cong \sqrt{p_N(1-p_N)/N} \quad . \tag{8}$$

In practice it might often be inconvenient to test the whole ensemble of N objects, and one may wish to test only $n \ll N$; in which case one may ask what is the error made in assigning the empirical value $p_n = q_1/n$ to the probability p, or p_N. The value p_n occurs with the probability

$$f(p_n) = C_n^{q_1} \cdot p^{q_1}(1-p)^{n-q_1} \tag{9}$$

and we may restrict to $n < \min(q, N-q)$, such that $0 \le q_1 \le n$. Introducing

$$J(\alpha) = \sum_{q_1=0}^{n} C_n^{q_1} \cdot (\alpha p)^{q_1}(1-p)^{n-q_1} = (1-p+\alpha p)^n \tag{10}$$

it is easy to show that the average empirical probability is p,

$$\overline{p_n} = \sum_{q_1=0}^{n} C_n^{q_1} \cdot \frac{q_1}{n} \cdot p^{q_1}(1-p)^{n-q_1} = \frac{1}{n} \cdot \frac{dJ}{d\alpha} |_{\alpha=1} = p \quad , \tag{11}$$

and its spread is given by

$$\begin{aligned} \overline{(\delta p_n)^2} &= \textstyle\sum_{q_1=0}^{n} C_n^{q_1} \cdot \left(\frac{q_1}{n} - p\right)^2 \cdot p^{q_1}(1-p)^{n-q_1} = \\ &= \tfrac{1}{n^2}\left(\tfrac{d^2J}{d\alpha^2} + \tfrac{dJ}{d\alpha}\right) |_{\alpha=1} - p^2 = \tfrac{p(1-p)}{n} \quad . \end{aligned} \tag{12}$$

For fixed n and N large enough one may say that p and p_N are given by the empirical probability p_n with an error $\sim \sqrt{p_N(1-p_N)/2n}$, which is less accurate than testing the whole ensemble by a factor $\sqrt{N/2n}$. For n large enough but still smaller than N one can say that the error made in attributing to p_N the value p_n is

$$\delta p_n \cong \sqrt{p_n(1-p_n)/2n} \quad . \tag{13}$$