THE TEACHING OF THE CALCULUS OF PROBABILITY

by

Dr. Arthur H. Copeland
Professor at the University of Michigan

TABLE OF CONTENTS

THE TEACHING OF THE CALCULUS OF PROBABILITY

There has always been considerable disagreement among experts concerning the significance of the relation between probabilities and the statistical data with which these probabilities are supposed to be somehow connected. Such controversy indicates both that the relation is important and that it is difficult to grasp. In presenting the theory of probability to students, one should therefore make every effort to clarify this relation. I shall give a brief outline of such a presentation.

Physical measurements. If we make n measurements $x^{(1)},x^{(2)},\ldots,x^{(n)}$ of some physical quantity these measurements will in general differ and it is natural to take the average $\frac{1}{n}\sum_{k=1}^{n}x^{(k)}$ as an estimate of the quantity measured. It is reasonable to suppose that in general this estimate improves as n increases. This situation can be formulated mathematically as follows: We imagine that an infinite sequence x of numbers $x^{(k)}$ is associated with the quantity to be measured. Thus

$$x = x^{(1)},x^{(2)},\ldots,x^{(k)},\ldots$$

We also imagine that there is associated with the above quantity a number p(x) which we can call the expected value of the quantity to be measured. We observe n terms of the sequence x by making n measurements. We assume that our estimate

$p_n(x) = \frac{1}{n}\sum_{k=1}^{n} x^{(k)}$ of the expected value tends to improve as n increases and that $\lim_{n \to \infty} p_n(x) = p(x)$. I shall have more to say concerning this assumption at a later point.

Functions of sequences. We are frequently confronted with the problem of estimating the expected value of some function of one or more physical quantities. Thus we introduce the following definition: Let f(u,v) be a function of the variables u and v and let

$$x = x^{(1)}, x^{(2)}, \ldots$$
and
$$y = y^{(1)}, y^{(2)}, \ldots$$

be two sequences. Then

$$f(x,y) = f(x^{(1)},y^{(1)}),\ f(x^{(2)},y^{(2)}), \ldots$$

Thus a function of two sequences is itself a sequence. For example

$$x + y = x^{(1)} + y^{(1)},\ x^{(2)} + y^{(2)}, \ldots$$

and from this equation it follows that

$$p(x+y) = p(x) + p(y).$$

The definition of a function of sequences is readily extended to the case of n sequences.

Constant sequences. We also have occasion to consider a function in which some of the arguments are sequences and the remaining arguments are parameters which do not vary from one measurement to the next. Thus we introduce the following definition: A sequence

$$a = a,\ a,\ a,\ \ldots$$

all of whose terms are the same is called a constant sequence or parameter. The same letter is used to denote both the sequence and the terms of the sequence. The ambiguity of notation does not seem to cause any difficulty.

Variance. As an example, the expression $p[(x-a)^2] = p(x^2) - 2ap(x) + a^2$ measures the average squared deviation from the parameter a. The minimum value of this expression with respect to a is $p(x^2) - p^2(x) = \sigma^2(x)$ and this minimum is attained when $a = p(x)$. The number $\sigma^2(x)$ is called the variance and its square root $\sigma(x)$ is called the standard deviation. In an analogous manner we can define the variance $\sigma_n^2(x) = p_n(x^2) - p_n^2(x)$ and the standard deviation $\sigma_n(x)$ for a finite sequence of measurements.

Events. If we let 1 and 0 denote respectively the success and failure of an event on a given trial, then an event may be regarded as a physical quantity the measurements of which can have only the values 1 and 0. If x is the corresponding sequence, then $p_n(x)$ is the success ratio for the first n trials and p(x) is the probability of the event.

Algebra of events. If x and y are 1,0-sequences, they can be regarded as the sequences of successes and failures associated with two events. It is easily seen that $\sim x = 1 - x$ represents the event "not x", $x \cdot y$ represents the event "x and y", and $x \vee y = \sim(\sim x \cdot \sim y) = x + y - x \cdot y$ represents the event "x or y (or both)". We have the relations

$$p(\sim x) = 1 - p(x), \; p(x \vee y) = p(x) + p(y) - p(x \cdot y).$$

The event "x if y" is denoted by $x \subset y$ and is defined as follows:

$$x \subset y = x^{(n_1)}, x^{(n_2)}, \ldots x^{(n_k)}, \ldots$$

where n_k is the trial on which the k-th success of y occurs. The sequence $x \subset y$ is in general infinite and is obtained as the result of a selection operation on the sequence x, the n-th term $x^{(n)}$ being selected if and only if the n-th trial of y is a success (i.e., $y^{(n)} = 1$). We have the relation

$$p(x)p(y \subset x) = p(x \cdot y).$$

Mutually exclusive events. Two events x and y are said to be mutually exclusive provided $x \cdot y = 0$. If the events* $x_1, x_2, \ldots, x_n$ are mutually exclusive, then

$$p(x_1 \vee x_2 \vee \ldots \vee x_n) = p(x_1) + p(x_2) + \cdots + p(x_n).$$

Independent events. The events $x_1, x_2, \ldots, x_n$ are said to be independent if

$$p(x_1 \cdot x_2 \cdot \ldots \cdot x_n) = p(x_1)p(x_2)\ldots p(x_n)$$

and if a similar condition holds for every subset of these events. If two events x and y are independent, then

$$p(x \subset y) = p(x) \quad \text{and} \quad p(y \subset x) = p(y).$$

This algebra forms the basis for the solution of the usual probability problems.

Fundamental function. The function $\varphi_1(u)$ depends on the variable u and the interval I, and is defined as follows

$$\varphi_I(u) = \begin{cases} 1 \text{ if u is in I} \\ 0 \text{ otherwise.} \end{cases}$$

If x is a sequence associated with a physical quantity, then $\varphi_1(x)$ is a 1,0-sequence representing an event which succeeds or fails on its k-th trial according as the k-th measurement of x does or does not fall within I. The probability that the measurement will fall within I is $p[\varphi_I(x)]$.

* Note that subscripts are used to indicate different sequences whereas superscripts are used to indicate the different terms of a given sequence.

Tchebycheff's inequality. Let I be the interval $p(x) - \varepsilon < u < p(x) + \varepsilon$. Then $\sim\varphi_I(u) \leqq [u-p(x)]^2/\varepsilon^2$ and hence we obtain the following inequality of Tchebycheff:

$$p[\sim\varphi_I(x)] \leqq \sigma^2(x)/\varepsilon^2.$$

Thus the probability that a measurement will be in error by more than ε (i.e., fall outside the interval) is at most $\sigma^2(x)/\varepsilon^2$.

Independence of physical quantities. Let x and y be the sequences associated with two physical quantities. Then x and y are independent provided the events $\varphi_I(x)$ and $\varphi_J(y)$ are independent for every pair of intervals I and J. It can be proved that if x and y are independent, then $p(x\cdot y) = p(x)p(y)$. If x and y are dependent, then

$$r(x,y) = [p(x\cdot y) - p(x)p(y)]/\sigma(x)\sigma(y)$$

measures the correlation between them. If $x_1,x_2,\ldots,x_n$ are independent, then

$$\sigma^2[(x_1 + x_2 + \cdots + x_n)/n] = [\sigma^2(x_1) + \sigma^2(x_2) + \cdots + \sigma^2(x_n)]/n^2.$$

Thus for the average $(x_1 + x_2 + \cdots + x_n)/n$ of n physical quantities, Tchebycheff's inequality takes the form

$$p\{\sim\varphi_I[(x + x_2 + \cdots + x_n)/n]\} \leqq [\sigma^2(x) + \sigma^2(x_2) + \cdots + \sigma^2(x_n)]/n^2\varepsilon^2.$$

Independence of individual observations. We shall introduce certain 1,0-sequences in terms of which we can give a precise meaning to the intuitive concept of independence of observations. Let (r,n) be a sequence in which the 1's occur in the terms whose superscripts are r + 1, r + n + 1, r + 2n + 1,... and the 0's occur in the remaining terms. Then the sequence*

* Note that in the definition of $x \subset y$, x need not be a 1,0-sequence whereas y must be such a sequence.

$x \subset (r,n)$ is obtained by selecting the $(r+1)$-st term of the sequence x and every n-th term thereafter. Note that the first terms of the sequences $x \subset (0,n)$, $x \subset (1,n), \dots, x \subset (n-1,n)$ constitute the first n terms of the sequence x (i.e., the first n observations). The second terms of these sequences constitute the second group of n terms of x, etc. We shall say that the observations are independent provided the sequences $x \subset (o,n)$, $x \subset (1,n), \dots, x \subset (n-1,n)$ are independent for every n. If this condition is satisfied and if further $p\{\varphi_I[x \subset (r,n)]\} = p[\varphi_I(x)]$ for every interval I and every pair of integers r,n such that $0 \leqq r < n$, then x is said to be admissible. It can be proved that if x is admissible, then $p[x \subset (r,n)] = p(x)$ and $\sigma[x \subset (r,n)] = \sigma(x)$. It is reasonable to assume that the sequence x associated with any physical quantity is admissible.

Error of the average of n trials. Let us consider the average X_n/n where

$$X_n = x \subset (0,n) + x \subset (1,n) + \cdots + x \subset (n-1,n)$$

and where x is admissible. The following formulas are readily established

$$p(X_n/n) = p(x), \quad \sigma^2(X_n/n) = \sigma^2(x)/n,$$

$$X_n/n = p_n(x),\ p_n[x \subset (n,1)],\ p_n[x \subset (2n,1)], \dots$$

Applying the Tchebycheff inequality we get

$$p[\sim \varphi_I(X_n/n)] \leqq \sigma^2(x)/n\varepsilon^2.$$

Thus the probability that a term of X_n/n shall be in error by more than ε is at most $\sigma^2(x)/n\varepsilon^2$. This is a precise formulation of a more common elliptical statement, namely, that the probability that the first term $p_n(x)$ shall be in error by more than ε is at most $\sigma^2(x)/n\varepsilon^2$. It is customary to replace $\sigma^2(x)$ by $\sigma_n^2(x)$ since the latter quantity can be calculated from the first n measurements. As the inequality is exceedingly generous it probably continues to hold after this replacement.

Distribution functions. Let I_s and I_s' be respectively the intervals $-\infty < u \leqq s$, $-\infty < u < s$ and let $p[\varphi_{I_s}(x)] = F(s+0)$, $p[\varphi_{I_s'}(x)] = F(s-0)$, and $[F(s-0) + F(s+0)]/2 = F(s)$. In general,[*] $F(s+0) = \lim_{\varepsilon\to 0} F(s+\varepsilon)$, $F(s-0) = \lim_{\varepsilon\to 0} F(s-\varepsilon)$, $F(s)$ is monotone, $F(-\infty) = 0$, and $F(+\infty) = 1$. Finally, if $I_{a,b}$ is the interval $a < u \leqq b$, then

$$p[\varphi_{I_{a,b}}(x)] = F(b+0) - F(a+0).$$

The function $F(s)$ is called the distribution function associated with the sequence x.

Integrals. Let $g(u)$ be a continuous function and let

$$G(s) = p[g(x)\varphi_{I_s}(x)].$$

Then it is easy to see that

$$m[F(b+0) - F(a+0)] \leqq G(b) - G(a) \leqq M[F(b+0) - F(a+0)],$$

where m and M are respectively the minimum and the maximum of $g(u)$ in $I_{a,b}$. Next suppose that $F(s)$ possesses a continuous derivative $F'(s)$. It follows from the above mean value

[*] It is possible to construct a sequence violating all of these conditions except the monotoneity but such sequences are exceptional and do not correspond to physical measurements.

property that $G'(s) = g(s)F'(s)$ and $G(s) = \int_{-\infty}^{s} g(t)F'(t)dt$. Hence

$$p[g(x)] = \int_{-\infty}^{+\infty} g(s)F'(s)ds = \int_{-\infty}^{+\infty} g(s)dF(s).$$

The expected value $p[g(x)]$ exists under much more general conditions than those stated above and can be taken as a definition of the Stieltjes integral on the right when the Riemann integral in the middle fails to exist. For example, let $x = x_1 + 2x_2 + \cdots + 6x_6$ where $x_1, x_2, \ldots, x_6$ are mutually exclusive 1,0-sequences such that $p(x_1) = p(x_2) = \ldots = p(x_6) = 1/6$. This sequence represents the throwing of a die. The distribution function $F(s)$ is discontinuous at the points $s = 1, 2, \ldots 6$ and constant elsewhere. We get

$$\int_{-\infty}^{+\infty} s dF(s) = p(x) = (1 + 2 + \cdots + 6)/6 = 3.5$$

and

$$\int_{-\infty}^{+\infty} s^2 dF(s) = p(x^2) = (1 + 4 + \cdots + 36)/6 = 91/6.$$

As the student is usually unfamiliar with Stieltjes integration, this approach is decidedly advantageous.

Moment generating functions. The function $p(e^{ixt}) = \varphi(t) = \int_{-\infty}^{+\infty} e^{ist}dF(s)$ (where t is a constant sequence) is called the moment generating function of the sequence x. It uniquely determines the distribution function as the following computation shows. First consider the integral

$$\frac{1}{2\pi i} \int_{-\infty}^{+\infty} \frac{e^{it} - e^{ixt}e^{-ist}}{t} dt = \begin{cases} 1 & \text{if } x < s \\ 1/2 & \text{if } x = s \\ 0 & \text{if } x > s \end{cases} = \frac{\varphi_{I_s'}(x) + \varphi_{I_s}(x)}{2}$$

Thus

$$p[\frac{1}{2\pi i} \int_{-\infty}^{+\infty} \frac{e^{it} - e^{ixt}e^{-ist}}{t} dt] = p[\frac{\varphi_{I_s'}(x) + \varphi_{I_s}(x)}{2}] = F(s).$$

But

$$p[\frac{1}{2\pi i}\int_{-\infty}^{+\infty}\frac{e^{it} - e^{ixt}e^{-ist}}{t}\,dt] = \frac{1}{2\pi i}\int_{-\infty}^{+\infty} p[\frac{e^{it} - e^{ixt}e^{-ist}}{t}]dt$$

$$= \frac{1}{2\pi i}\int_{-\infty}^{+\infty}\frac{e^{it} - p(e^{ixt})e^{-ist}}{t}\,dt.$$

Hence

$$F(s) = \frac{1}{2\pi i}\int_{-\infty}^{+\infty}\frac{e^{it} - \varphi(t)e^{-ist}}{t}\,dt.$$

The probability integral. Let $x_k = x \subset (k+1,n)$ and $X_n = x_1 + x_2 + \cdots + x_n$, where x is admissible, $p(x) = 0$, and $p(x^2) = 1 = \sigma^2(x)$. Then

$$p(e^{iX_nt/\sqrt{n}}) = p(e^{ix_1t/\sqrt{n}})p(e^{ix_2t/\sqrt{n}})\ldots p(e^{ix_nt/\sqrt{n}})$$

$$= p^n(e^{ixt/\sqrt{n}}) = p^n(1 + \frac{ixt}{\sqrt{n}} - \frac{x^2t^2}{2n} - \ldots) = (1 - \frac{t^2}{2n} - \ldots)^n.$$

Hence

$$\lim_{n\to\infty} p(e^{iX_nt/\sqrt{n}}) = \lim_{n\to\infty}(1 - \frac{t^2}{2n} - \ldots)^n = e^{-t^2/2}.$$

Thus if $F_n(s)$ is the distribution function for the sequence $X_n/\sqrt{n}$, it follows that

$$\lim_{n\to\infty} F_n(s) = \frac{1}{2\pi i}\int_{-\infty}^{+\infty}\frac{e^{it} - e^{-t^2/2}e^{-ist}}{t}\,dt = \Phi(s)$$

where

$$\Phi(s) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{s} e^{-t^2/2}\,dt.$$

The latter equality can be established by showing that

$$\int_{-\infty}^{+\infty} e^{ist}d\,\Phi(s) = e^{-t^2/2}.$$

The problem of application.[*] In conclusion, I should like to make a remark about the problem of application. Suppose we make the physical hypothesis that the probability of some event is 1/2, i.e., that the associated sequence x is such that $p(x) = 1/2$. Suppose further that we make 1000 trials of the event and obtain the result $p_{1000}(x) = .491$. We are tempted to believe that the result is a confirmation of our hypothesis. However, if the result confirms the hypothesis, it must verify some implication of the hypothesis. But our hypothesis actually does not imply anything about the value of $p_{1000}(x)$. No one of the possible values 0, .001, .002, ... 1 is excluded by the hypothesis $p(x) = 1/2$. This argument would seem to indicate that there is no way of verifying any probability (or any expected value) by means of statistical data. However, this is not the case. In fact, *the hypothesis* $p(x) = w$ *has an infinitude of verifiable implications*. Each implication depends on a pair of positive numbers ε, N and is stated as follows: *There exists an integer* n *such that* $n > N$ *and* $|p_n(x) - w| < \varepsilon$. For example, the hypothesis $p(x) = 1/2$ implies the existence of an integer n such that $n > 500$ and $|p_n(x) - 1/2| < .01$. Since we have obtained the value $p_{1000}(x) = .491$, the number $n = 1000$ satisfies the conditions and therefore the hypothesis is verified by this result.

Of course, this theory leaves much to be desired. Physical experiment cannot establish either the truth or the falsity of a hypothesis concerning a probability. However,

[*] See also the editor's note following this section concerning the same problem.

suppose that someone makes a hypothesis concerning a probability $p(x)$ and that he makes the claim that there is no one of the implications of this hypothesis which cannot be verified experimentally. He is then claiming that $p(x)$ is a limit point of the sequence $p_n(x)$. If further the above individual subscribes to the physical hypothesis that such limit points are unique, he thereby assigns a physical meaning to $p(x)$. Undoubtedly, practical objections can be raised. It may happen that verification of a given implication is impossible within the lifetime of the experimenter or of the object on which experimentation is performed. Thus in spite of the fact that we have clarified the relation between probability and experiment, some vagueness remains. In discussing this subject, Wald (On the principles of statistical inference, Notre Dame Mathematical Lectures, No. 1) remarks that "... such vagueness is always associated with the application of theory to real phenomena."

In this outline I have indicated some of the ways in which probabilities can be computed from expected values and from other probabilities. I have also given a brief explanation of the precise sense in which these computations can be verified by statistical data. I hope this will help clarify the relations between probabilities and physical measurement.