AN EXACT TEST FOR RANDOMNESS IN THE NON-PARAMETRIC CASE BASED ON SERIAL CORRELATION[1]

By A. Wald and J. Wolfowitz

Columbia University

1. Introduction. A sequence of variates $x_1, \cdots, x_N$ is said to be a random series, or to satisfy the condition of randomness, if $x_1, \cdots, x_N$ are independently distributed with the same distribution; i.e., if the joint cumulative distribution function (c.d.f.) of $x_1, \cdots, x_N$ is given by the product $F(x_1) \cdots F(x_N)$ where $F(x)$ may be any c.d.f.

The problem of testing randomness arises frequently in quality control of manufactured products. Suppose that x in some quality character of a product and that $x_1, x_2, \cdots, x_N$ are the values of x for N consecutive units of the product arranged in some order (usually in the order they were produced). The production process is said to be in a state of statistical control if the sequence $(x_1, \cdots, x_N)$ satisfies the condition of randomness. A number of tests of randomness have been devised for purposes of quality control, all having the following features in common: 1) They are based on runs in the sequence $x_1, \cdots, x_N$. 2) The test procedure is invariant under topologic transformation of the x-axis, i.e., the test procedure leads to the same result if the original variates $x_1, \cdots, x_N$ are replaced by $x'_1, \cdots, x'_N$ where $x'_\alpha = f(x_\alpha)$ and $f(t)$ is any continuous and strictly monotonic function of t. 3) The size of the critical region, i.e., the probability of rejecting the hypothesis of randomness when it is true, does not depend on the common c.d.f. $F(x)$ of the variates $x_1, \cdots, x_N$. Condition (3) is *a fortiori* fulfilled if condition (2) is satisfied and if $F(x)$ is continuous. The fulfillment of condition (3) is very desirable, since in many practical applications the form of the c.d.f. $F(x)$ is unknown.

Tests of randomness are of importance also in the analysis of time series (particularly of economic time series) where they are frequently based on the so-called serial correlation. The serial correlation coefficient with lag h is defined by the expression[2] (see, for instance, Anderson [1])

$$(1) \qquad R_h = \frac{\sum_{\alpha=1}^{N} x_\alpha x_{h+\alpha} - \left(\sum_{\alpha=1}^{N} x_\alpha\right)^2 \Big/ N}{\sum_{\alpha=1}^{N} x_\alpha^2 - \left(\sum_{\alpha=1}^{N} x_\alpha\right)^2 \Big/ N}$$

where $x_{h+\alpha}$ is to be replaced by $x_{h+\alpha-N}$ for all values of α for which $h + \alpha > N$. The distribution of R_h has recently been studied by R. L. Anderson [1], T. Koopmans [2], L. C. Young [3], J. v. Neumann [4, 5], B. I. Hart and J. v. Neu-

[1] Presented to the Institute of Mathematical Statistics and the American Mathematical Society at a joint meeting at New Brunswick, New Jersey, on September 13, 1943.

[2] Some authors (see, for instance, [2] p. 27, equation (61)) use a non-circular definition.