PSYCHOMETRIKA—VOL. 36, NO. 1
MARCH, 1971

BIAS AND RUNS IN DICE THROWING AND RECORDING: A FEW MILLION THROWS*

GUDMUND R. IVERSEN, WILLARD H. LONGCOR†, FREDERICK MOSTELLER, JOHN P. GILBERT, AND CLEO YOUTZ

An experimenter threw individually 219 different dice of four different brands and recorded even and odd outcomes for one block of 20,000 trials for each die—4,380,000 throws in all. The resulting data on runs offer a basis for comparing the observed properties of such a physical randomizing process with theory and with simulations based on pseudo-random numbers and RAND Corporation random numbers. Although generally the results are close to those forecast by theory, some notable exceptions raise questions about the surprise value that should be associated with occurrences two standard deviations from the mean. These data suggest that the usual significance level may well actually be running from 7 to 15 percent instead of the theoretical 5 percent.

The data base is the largest of its kind. A set generated by one brand of dice contains 2,000,000 bits and is the first handmade empirical data of such size to fail to show a significant departure from ideal theory in either location or scale.

1. Introduction

How well do the laws of chance actually work? When a die is repeatedly thrown and its outcomes recorded, do imperfections in the die, in the throwing, in the perception of the outcome, and in recording appear? What sorts of deviations from chance do we find?

Weldon's dice data [Fry, 1965] and Kerrich's coin tossing monograph [Kerrich, 1946] both give us some experience with large bodies of data produced by humanly run physical randomizing devices whose idealized probabilities and properties are known to a good approximation. In a sense, such experiments are controls on other experiments where probability plays an important role. For example, such dice and coin experiments give us an idea of how seriously we should take small departures from mathematically predicted results in investigations where we search for small departures from a standard. They do this by showing the sizes and kinds of departures observed in an experiment with *no planned* human or material effects. They are placebo experiments. If one does not believe in extra-sensory perception, then many ESP investigations also would be judged to qualify, but if one

* The analysis was facilitated by a National Science Foundation grant GS-341 and and its continuation GS-2044X. It forms part of a larger study of data analysis.

† Mr. Longcor is from Waukegan, Illinois; the other authors are from Harvard University. Dr. Iversen has moved to the University of Michigan.

does believe in ESP then, in such experiments, departures from mathematical forecasts are contaminated by small effects over and above the procedural ones mentioned above. In the latter case, results of experiments like the one reported here are especially relevant because they provide a baseline for departures from perfection (unless of course the experimenter is making use of psychokinesis).

Examining such data gives us a background of experience with physical devices for carrying out randomization. This experience can be compared with the results of pseudo-randomizing devices such as those used in high-speed computers.

We report here an analysis of long sequences of throws of dice of four brands by a single experimenter, together with two control series—one based on the RAND random digits [1955], the other produced by a pseudo-random number generator on a high-speed computer.

In the present investigation the experimenter, Willard Longcor, was curious as to whether enormously long sequences of throws would continue to behave according to the laws of chance. He had taken dice throwing and recording as a long-time hobby. In advance of the investigation but after pilot work, he and Frederick Mosteller agreed that the outcome of the throw of a single die would be recorded as even or odd, because Longcor had much experience with this particular way of recording. Testing for bias in dice with holes for pips might have been more powerful using "low" (1, 2, 3) versus "high" (4, 5, 6), but in pilot work Longcor found that this method produced recording errors for him and he scrapped the data based on it and returned to the more familiar even and odd. Recording the actual number was not an option the experimenter wished to try at that time. The matter is discussed further in Appendix 2.

Longcor and Mosteller agreed that lengths of runs of evens would be the basis for summarizing the data. Inevitably this focuses the experimenter's attention on runs and their lengths, but perhaps not more than it would naturally have been. Long runs are the natural source of surprise. Therefore anyone would look for bias in them.

We shall then be attending to long runs, but we also want to look at the joint behavior of numbers of runs of various lengths, through their covariances, and at other aspects of the data, because biases might very easily appear in these second-order statistics that few can know or compute.

Some sorts of bias, of course, cannot be detected by this even-odd method of recording. For example, if the probabilities of opposite sides are equally inflated and other opposites equally deflated as they might be if the die were a rectangular parallelopiped instead of a cube, this bias would not be detected by our recording method. And more generally, any bias leaving $P(\text{even}) = P(\text{odd})$ would not be detected.

The distribution theory of runs of two or more kinds of elements was