Vol. 16, No. 10. October, 1919

THE

PSYCHOLOGICAL BULLETIN

MATHEMATICAL VS. SCIENTIFIC SIGNIFICANCE

BY EDWIN G. BORING

Clark University

The conclusions attained in researches involving measuremen are very frequently based on estimates of identity or difference between two or more characters measured; the scientist finds that he has to deal with differences between two constants or with the divergence of one function from another.[1] Every difference or every measure of disparitv that thus comes under consideration is necessarily obtained from given data with a certain measure of precision (*e.g.*, a P.E.) that gives rise to the question whether the difference is "significant" or not. When a difference is large with respect to its probable error, for example, it is assumed to be "significant," and when it is small with respect to its probable error it is said to be "less significant" or "insignificant." For this reason it is customary in the case of a simple difference between the means of two arrays to measure the "significance" of the difference by its ratio to its P.E.

It is usually held that this ratio is not a direct measure of significance, and needs to be translated into a scale of the 'probability of difference' by the use of a table of the probability integral. The values from such a table give us a series of numbers ranging from zero, when the difference is infinitesimal with respect to its P.E. (not "significant"), to unity, when the difference is infinite with respect to its P.E. ("significant"). This scale of probabilities is sometimes spoken of as the "probability that the difference is

[1] See my discussion with S. W. Fernberger: *Amer. J. of Psychol.*, 1916, **27**, 315–319; 1917, **28**, 454–459; PSYCHOL. BULL. 1917, **14**, 110–113; also *Tables for Statisticians and Biometricians*, ed. by K. PEARSON, 1914, pp. xvii f.

not due to chance." It is also used as a measure of "homogeneity" and "heterogeneity"; for, if the difference between two samples of the same data is "insignificant," the data may be thought to be "homogeneous," whereas, if the difference is large and therefore "significant," the data from the two samples taken together may be considered "heterogeneous."[1]

Pearson applies a similar principle in his "measure of the goodness of fit" between two curves. With his procedure one obtains a value χ^2 from summing the differences between two curves and, taking into account the number of cases, makes use of a table to determine the probability that the deviation of the one curve from the other is merely "random."[2]

It is a common experience of scientific persons working with human data that these formulæ frequently give values for the probability of differences that are "too high." One works, for example, with the performances of a group of women and a group of men in a mental test and one finds a "significant" difference—perhaps a probability that 99 times out of 100 the men will do better than the women,—and yet one is convinced that there is no "truly significant" difference indicated. Or one determines the deviation of an observed curve from an ideal form and finds, let us say, that only 2 times in 100 would data that tend to follow the ideal form deviate as much from the ideal as do the observed data; and yet in plotting the observations along with the theoretical form he may note that the two functions are sensibly the same, and may feel inclined (if he is not scared off by Mr. Pearson) to say that the ideal function actually does represent his data. It is with the basis for this particular scientific attitude that I am concerned.[3]

It appears that the apparent inconsistency between scientific intuition and mathematical result is not due to the unreliability of professional opinion, but to the fact that scientific generalization is a broader question than mathematical description. In scientific work we deal with samples, whereas we are always interested in the

[1] *Cf. opp. citt.*, especially *Amer. J. of Psychol.*, 28, 451 ff., and V. Henri, *L'Année psychol.*, 1898, 5, 153 ff.

[2] K. Pearson, *Phil. Mag.*, 1900, 50, 157–175; W. P. Elderton, *Biometrika*, 1902, 1, 155–163; *Tables for Statisticians*, *op. cit.*, pp. xxxi ff., 26 ff.

[3] The differences between Pearson on the one hand, and Merriman and Airy on the other, *Phil. Mag.*, *op. cit.*, 171 ff. are of this order; Pearson is statistician, and Merriman and Airy scientists. It is interesting to find Pearson shifting ground in *Biometrika*, 2, 1903, p. 367, from a statistical result that lengths of forearm do not fit the Gaussian curve to a scientific conclusion that they do.

larger groups of which the samples are intended to be representative. The mathematical formulae do truly measure the difference between the particular samples observed. Whenever we can assume that these samples "truly" represent the total group, then the mathematical method also indicates the probability of a difference between the groups represented. A sample "truly" represents a group when the mode of variation within the sample is the same as the variation within the group at large: this is what is meant when we say we have an "unselected" sample. But anyone who has attempted to obtain "unselected" samples with human material knows what very careful selection is required to achieve this "unselected" state.[1] There are many uncontrollable factors that enter into the getting of human stuff; human beings are usually resistent to an indiscriminate mixing-up and to that arbitrary selection combined with complete ignorance of the nature of the individuals involved which constitutes "chance selection." So it happens that the competent scientist does the best he can in obtaining unselected samples, makes his observations, computes a difference and its "significance," and then—absurd as it may seem—very often discards his mathematical result, because in his judgment the mathematically "significant" difference is nevertheless not large compared with what he believes is the discrepancy between his samples and the larger groups which they represent.

It is useless to try to limit the scientist to the mere description of his samples. Science begins with description but it ends in generalization. And, since in the nature of the case it is impossible for him to state in numerical terms the degree of representativeness that his samples possess, conclusions must ultimately be left to the scientific intuition of the experimenter and his public. Such an outcome with respect to the measure of the probability of difference is not wholly satisfactory but it is inescapable. It is equivalent merely to saying that, given only approximate control of experimental conditions, only approximate results can be achieved. A knowledge of the "probability that a difference is not due to chance" is distinctly worth while on the descriptive side; but this measure of significance does not necessarily apply to the general class for which a sample stands. In certain cases it may so apply, but ordi-

[1] Statisticians' rules for obtaining "chance conditions" and "random samples," though generally failing of an appreciation of the logical truth that complete ignorance is the sole condition of chance, show how hard this particular kind of ignorance is to achieve.

narily there is a constant factor operative in the selection of human material which must be taken into account and which frequently offsets a demonstrably "significant" difference that has been made out between the samples. It is for this reason that mathematical measures of difference are apt to be too high and may need to be discounted in arriving at a scientific conclusion. The case is one of many where statistical ability, divorced from a scientific intimacy with the fundamental observations, leads nowhere.

AN OBSERVATION OF THE PURKINJE PHENOMENON IN SUB-TROPICAL MOONLIGHT

BY STEPHEN G. RICH

Mansfield Park School, Durban, South Africa

The observation here reported was made at East London, South Africa, July 12, 1919. The town is in 32° S. Latitude, at the extreme southern end of the eastern sub-tropical margin of Africa. The moon was on this occasion approximately full; and at 8:30 P.M., when the observation was made, was at the zenith. The sky was cloudless, and the place of observation was beyond the area served by street-lights. The colors of objects seen by moonlight were verified by a daylight visit to the same places on the following morning.

In full moonlight a limited range of colors was visible. Reds were especially noticeable. A brick wall showed its characteristic hue; orange-red tiled roofs were plainly recognizable in color; maroon-red painted roofs were seen as a very deep brownish-red; a bright carmine letter-box appeared a dull crimson; but a dark brown-red painted roof appeared black. Greens were not all recognizable. Pine-trees appeared black; pepper-trees (*Schinus molle*), greenish-gray in daylight, were gray; but hibiscus and aloe leaves were noticeably green. My dark blue suit appeared black; the hue was identical with that of my fountain-pen held against the cloth. Yellows and violets were not observed.

In the shadow of a pine-tree all hues save a maroon-red, which became a deep brownish-red, vanished. A pair of tan shoes which I was wearing were of a dark russet color. This was clearly visible in the direct moonlight, but vanished utterly in the shadow, becoming a medium gray.