The Meaning of a Significance Level
Author(s): G. A. Barnard
Source: *Biometrika*, Vol. 34, No. 1/2 (Jan., 1947), pp. 179-182
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: http://www.jstor.org/stable/2332521
Accessed: 23-09-2017 15:03 UTC

THE MEANING OF A SIGNIFICANCE LEVEL

By G. A. BARNARD

A level of significance is a probability. To say that a given result is significant on the 5 % level means that some class of events has probability 0·05. Now whatever theory we may hold as to the nature of probability, in order to give a statement of probability a precise meaning we must refer to some reference class, or set of data, on which the probability is calculated. What is the reference class involved in a level of significance?

To many people the answer to this question seems simple enough. The reference class involved is the set of indefinite (possibly imaginary) repetitions of the experiment which gave the result in question. Otherwise put, the data, on which the probability is calculated, are the external conditions of the experiment. The following example indicates, however, that the meaning of this reference class is not always clear. The example is a modified form of one given by Prof. R. A. Fisher in a letter to the author.

Suppose we have a bag of chrysanthemum seeds, known to give plants having white flowers or plants having purple flowers, no other colours being possible. We suspect that the proportions of white and purple seeds are equal, and to test this hypothesis we select at random ten seeds from the bag, and plant them. Nine of the plants grow to maturity, and all of them have white flowers. On what level of significance can we reject the hypothesis of equality of proportions? We may assume that white and purple plants are equally viable.

It would be natural to argue that, if white and purple flowers were equally likely, the probability of our result would be $1/2^9$. If there is no reason to suspect an excess of white rather than an excess of purple flowers, we must add to this the probability of getting nine purple flowers, which is also $1/2^9$, giving a total probability of $1/2^8$. The hypothesis of equality of proportions would then be rejected on the 1/256, or the 0·3906 % level of significance. But if we did this our reference class would not be the set of indefinite repetitions of the experiment, in its ordinary meaning.

A repetition of the experiment, in its ordinary meaning, would consist of another selection of ten seeds from the bag, and their planting and growth. On such another occasion all ten plants might grow to maturity, or all or some might die. These possibilities have not been taken into account in our calculation of probability, so far.

To allow for the possible variation in the number of plants which grow, we might lay out the set of all possible results of the experiment as in Fig. 1, where n denotes the number of plants that grow, and r denotes the excess of white over purple. Thus any point in the figure can be referred to uniquely by its co-ordinates (n, r). If we now introduce a parameter p, to denote the probability (if it exists) that a plant will grow to maturity, given that it has been selected, the probability associated with the point (n, r) on the hypothesis of equality of proportions of white and purple will be

$$W(n,r;p) = \frac{10!}{n!(10-n)!}p^n(1-p)^{10-n}\frac{n!\,2^{-n}}{(\frac{1}{2}(n+r))!\,(\frac{1}{2}(n-r))!},$$

and since this is a function of the unknown p, we have a special problem of arranging the points (n, r) in order of significance before we can establish a test. The situation in this respect is similar to that dealt with in the paper on 2 × 2 tables, printed earlier in this issue (Barnard, 1946, pp. 123–38 above).

Proceeding as in the earlier paper, we notice first that the same level of significance must apply to (n, r) as to $(n, -r)$, so that we can confine our further considerations to the upper half of the diagram. Now in this half, the transition from (n, r) to $(n+1, r+1)$ means we discover that one of the plants which failed to grow in our case, was in fact a white-flowered plant. In this case our conviction that there is an excess of white-flowered plants would be strengthened, so that $(n+1, r+1)$ would be reckoned more significant than (n, r). Similarly, going from (n, r) to $(n+1, r-1)$ would mean that a missing plant was found to be purple, and this would weaken our belief in an excess of white-flowered plants; consequently, (n, r) would be reckoned more significant than $(n+1, r-1)$. Finally, going from (n, r) to $(n+2, r)$ would mean growing two more plants, one purple and one white, and this would increase our tendency to believe in the equality of proportions. Consequently, (n, r) would be reckoned more significant than $(n+2, r)$. These principles taken together imply that points lying north-east, or west, of a given point (n, r), or between these two directions, would be reckoned more significant than (n, r); while, conversely, points lying east to south-west (inclusive) from (n, r) would be reckoned less significant than (n, r). The relative significance of points lying inside the half-quadrants north-east to east and south-west to west would remain undetermined.

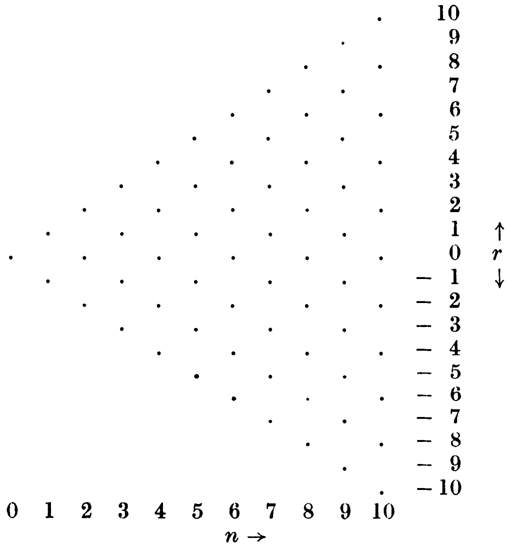

Fig. 1

We could now proceed as in the paper (1), building up a test, consistent with the above partial ordering, in such a way as to make the significance or otherwise of our result depend as little as possible on any knowledge we may have about the value of p. But we need not carry this through for the result we have quoted, since our conditions by themselves require that the only points in the diagram which should be reckoned not less significant than our result are the points (9, 9), (9, −9), (10, 10) and (10, −10). The probability associated with these four points is

$$P(9, 9; p) = 2(10p^9(1-p).2^{-9}+p^{10}2^{-10})$$
$$= (p/2)^9(20-19p),$$

the maximum value of which occurs when $p = 18/19$, and is $P_m(9, 9) = 0{\cdot}002413$. Thus on this basis we should conclude that our result was significant on the 0·2413 % level.

The difference between the first result, 0·3906 %, and the second, 0·2413 %, is in practice negligible. Somewhat larger differences will be found in other similar cases, however, and it seems worth while to try to clarify the cause of the discrepancy.

Consider three possible causes for the failure of the tenth plant to grow to maturity:

(1) The bag from which the seed was taken is known to contain a proportion of dead seeds, which are physically indistinguishable from the live ones, and the tenth seed planted happened to be one of these. The conditions of growth were such that any live seed planted would have grown.

(2) The tenth plant happened to be attacked by a soil pest, which destroyed it.

(3) The statistician trod on the tenth plant while running for a bus; otherwise, it would have grown.

If we now consider what would happen in these three cases if the experiment were repeated, in case (1) we should be just as uncertain as before how many plants would grow, out of those selected. In case (2), we might or might not happen to strike a good year for the pest in question, so that we might or might not have a similar accident recurring. In case (3) we should obviously give the statistician firm instructions not to be careless, and then we could be reasonably certain that all the plants selected would grow.*

In the first case, we can suppose that the proportions of white, purple, and dead seeds in the bag are, respectively, p_1, p_2, and $1-(p_1+p_2)$; and the purpose of our experiment is to test the hypothesis $p_1 = p_2$. In this case, putting $p_1+p_2 = p$, we can clearly apply the analysis of Fig. 1, and the appropriate level of significance is 0·2413 %.

In the third case, the situation actually realized is just what it would have been if we had warned the statistician beforehand, and then thrown one of the ten seeds back into the bag. Thus our effective sample size here is 9, and the appropriate level of significance is 0·3906 %.

In the second case, the answer depends on our attitude to the set of accidents of which the pest is a specimen. If this set of accidents is regarded as a stable set of chance causes we may be justified in representing its effect on the growth of our plants by the probability p. If, on the other hand, the incidence of such pests undergoes, say, regular cyclical fluctuations from year to year, so that its incidence is to some extent predictable, if not wholly controllable, then we should not be justified in assuming the existence of a real probability corresponding to our parameter p. We should, to be on the safe side, in this case allow for the possibility that experimental technique might improve in the future, to such an extent as to eliminate the possibility of such accidents. Thus, adopting this conservative attitude to our results, we should here treat the effective sample size as 9. The repetitions of the experiment which we have in mind would then be imaginary repetitions, in which experimental technique was supposed to be better than it is now, and we have as much control over pests as we have over statisticians.

The general situation illustrated by this example can be described in terms of the notion of 'isolate' introduced by Prof. H. Levy (1931). In making an experiment, we try to construct an isolate—a system, or part of the world, which we suppose has relatively little interaction with the rest of the world, and which, for practical purposes, may be considered on its own. This isolate may contain within itself all the systems of chance causes which are

* It is not suggested that the three cases exhaust the multiplicity of types which might arise in practice. As Prof. Pearson has pointed out, if it were not the statistician, but his three-year-old son who was the vandal in case (3), we should have here a situation intermediate between our second and third instances.

regarded as affecting, to any practical extent, the results of the experiment. Such is the case in (1), where all the chance causes involved in the experiment are supposed given in the bag which is the subject of the experiment. Here, then, we are dealing with a 'good isolate', whose interaction with the rest of the world is really negligible, and chance causes operate within the isolate.

In case (3), on the other hand, we are dealing with an imperfect isolate. The outside world, in the shape of the statistician, interacts with our isolate to an extent not negligible in practice. Fortunately, in this case we are able to construct a smaller isolate, consisting of the nine surviving plants, in which the interactions with the outside world are negligible. In case (2), there may be some doubt as to what isolate we are discussing. If we regard soil pests and such things as included in the isolate, and represent them as a stable set of chance causes, then we are entitled to analyse as in case (1); but if the pests are not included in the isolate, we should analyse as in case (3).

Statistical tests are applicable to at least two types of experiment. First, to experiments in which the isolate studied contains within itself a system of chance causes which may influence the results. And second, to experiments in which the isolate studied is not a 'good' isolate, and the residual interactions with the rest of the world may affect the results. There may also be mixed cases.

The distinction between the two types may also be brought out in relation to the necessity or otherwise of an 'artificial' randomization procedure, using random digits or the like. In the first type, such an artificial randomization procedure is not strictly necessary; for example, with our bag of seeds, the bag itself, and its physically indistinguishable contents, forms a perfectly adequate randomizer. We have in this case, as it were, an impermeable shield around the system, which prevents any external shocks from affecting the system. In the second type of experiment, we need to ensure that the interactions with the outside world will not mask the results we are interested in; and if we cannot ensure a practically complete separation from the outside world, then the effect of external intereactions must be randomized, by a special procedure. The randomization here acts like a shock absorber, specially placed around the experiment to distribute external shocks evenly through the system.

In the first type of experiment, the reference class to which the significance level applies is in fact the set of indefinite repetitions of the experiment in question. In the second type of experiment, the reference class is an ideal set, in which the accidental influences of the outside world repeat themselves exactly, while the effect of these accidents on the system varies as a result of the special randomization.

REFERENCES

Barnard, G. A. (1946). Significance tests for 2 × 2 tables. *Biometrika*, **34**, 123.

Levy, H. (1931). *The Universe of Science*. London: Watts and Co.