2. Belief in the law of small numbers

Amos Tversky and Daniel Kahneman

"Suppose you have run an experiment on 20 subjects, and have obtained a significant result which confirms your theory ($z = 2.23$, $p < .05$, two-tailed). You now have cause to run an additional group of 10 subjects. What do you think the probability is that the results will be significant, by a one-tailed test, separately for this group?"

If you feel that the probability is somewhere around .85, you may be pleased to know that you belong to a majority group. Indeed, that was the median answer of two small groups who were kind enough to respond to a questionnaire distributed at meetings of the Mathematical Psychology Group and of the American Psychological Association.

On the other hand, if you feel that the probability is around .48, you belong to a minority. Only 9 of our 84 respondents gave answers between .40 and .60. However, .48 happens to be a much more reasonable estimate than .85.[1]

Apparently, most psychologists have an exaggerated belief in the likelihood of successfully replicating an obtained finding. The sources of such

[1] The required estimate can be interpreted in several ways. One possible approach is to follow common research practice, where a value obtained in one study is taken to define a plausible alternative to the null hypothesis. The probability requested in the question can then be interpreted as the power of the second test (i.e., the probability of obtaining a significant result in the second sample) against the alternative hypothesis defined by the result of the first sample. In the special case of a test of a mean with known variance, one would compute the power of the test against the hypothesis that the population mean equals the mean of the first sample. Since the size of the second sample is half that of the first, the computed probability of obtaining $z \geq 1.645$ is only .473. A theoretically more justifiable approach is to interpret the requested probability within a Bayesian framework and compute it relative to some appropriately selected prior distribution. Assuming a uniform prior, the desired posterior probability is .478. Clearly, if the prior distribution favors the null hypothesis, as is often the case, the posterior probability will be even smaller.

This chapter originally appeared in *Psychological Bulletin*, 1971, 2, 105–10.

beliefs, and their consequences for the conduct of scientific inquiry, are what this paper is about. Our thesis is that people have strong intuitions about random sampling; that these intuitions are wrong in fundamental respects; that these intuitions are shared by naive subjects and by trained scientists; and that they are applied with unfortunate consequences in the course of scientific inquiry.

We submit that people view a sample randomly drawn from a population as highly representative, that is, similar to the population in all essential characteristics. Consequently, they expect any two samples drawn from a particular population to be more similar to one another and to the population than sampling theory predicts, at least for small samples.

The tendency to regard a sample as a representation is manifest in a wide variety of situations. When subjects are instructed to generate a random sequence of hypothetical tosses of a fair coin, for example, they produce sequences where the proportion of heads in any short segment stays far closer to .50 than the laws of chance would predict (Tune, 1964). Thus, each segment of the response sequence is highly representative of the "fairness" of the coin. Similar effects are observed when subjects successively predict events in a randomly generated series, as in probability learning experiments (Estes, 1964) or in other sequential games of chance. Subjects act as if *every* segment of the random sequence must reflect the true proportion: if the sequence has strayed from the population proportion, a corrective bias in the other direction is expected. This has been called the gambler's fallacy.

The heart of the gambler's fallacy is a misconception of the fairness of the laws of chance. The gambler feels that the fairness of the coin entitles him to expect that any deviation in one direction will soon be cancelled by a corresponding deviation in the other. Even the fairest of coins, however, given the limitations of its memory and moral sense, cannot be as fair as the gambler expects it to be. This fallacy is not unique to gamblers. Consider the following example:

> The mean IQ of the population of eighth graders in a city is *known* to be 100. You have selected a random sample of 50 children for a study of educational achievements. The first child tested has an IQ of 150. What do you expect the mean IQ to be for the whole sample?

The correct answer is 101. A surprisingly large number of people believe that the expected IQ for the sample is still 100. This expectation can be justified only by the belief that a random process is self-correcting. Idioms such as "errors cancel each other out" reflect the image of an active self-correcting process. Some familiar processes in nature obey such laws: a deviation from a stable equilibrium produces a force that restores the equilibrium. The laws of chance, in contrast, do not work that way: deviations are not canceled as sampling proceeds, they are merely diluted.

Thus far, we have attempted to describe two related intuitions about chance. We proposed a representation hypothesis according to which people believe samples to be very similar to one another and to the population from which they are drawn. We also suggested that people believe sampling to be a self-correcting process. The two beliefs lead to the same consequences. Both generate expectations about characteristics of samples, and the variability of these expectations is less than the true variability, at least for small samples.

The law of large numbers guarantees that very large samples will indeed be highly representative of the population from which they are drawn. If, in addition, a self-corrective tendency is at work, then small samples should also be highly representative and similar to one another. People's intuitions about random sampling appear to satisfy the law of small numbers, which asserts that the law of large numbers applies to small numbers as well.

Consider a hypothetical scientist who lives by the law of small numbers. How would his belief affect his scientific work? Assume our scientist studies phenomena whose magnitude is small relative to uncontrolled variability, that is, the signal-to-noise ratio in the messages he receives from nature is low. Our scientist could be a meteorologist, a pharmacologist, or perhaps a psychologist.

If he believes in the law of small numbers, the scientist will have exaggerated confidence in the validity of conclusions based on small samples. To illustrate, suppose he is engaged in studying which of two toys infants will prefer to play with. Of the first five infants studied, four have shown a preference for the same toy. Many a psychologist will feel some confidence at this point, that the null hypothesis of no preference is false. Fortunately, such a conviction is not a sufficient condition for journal publication, although it may do for a book. By a quick computation, our psychologist will discover that the probability of a result as extreme as the one obtained is as high as $\frac{3}{8}$ under the null hypothesis.

To be sure, the application of statistical hypothesis testing to scientific inference is beset with serious difficulties. Nevertheless, the computation of significance levels (or likelihood ratios, as a Bayesian might prefer) forces the scientist to evaluate the obtained effect in terms of a *valid* estimate of sampling variance rather than in terms of his subjective biased estimate. Statistical tests, therefore, protect the scientific community against overly hasty rejections of the null hypothesis (i.e., Type I error) by policing its many members who would rather live by the law of small numbers. On the other hand, there are no comparable safeguards against the risk of failing to confirm a valid research hypothesis (i.e., Type II error).

Imagine a psychologist who studies the correlation between need for achievement and grades. When deciding on sample size, he may reason as follows: "What correlation do I expect? $r = .35$. What N do I need to make the result significant? (Looks at table.) $N = 33$. Fine, that's my sample."

The only flaw in this reasoning is that our psychologist has forgotten about sampling variation, possibly because he believes that any sample must be highly representative of its population. However, if his guess about the correlation in the population is correct, the correlation in the sample is about as likely to lie below or above .35. Hence, the likelihood of obtaining a significant result (i.e., the power of the test) for $N = 33$ is about .50.

In a detailed investigation of statistical power, J. Cohen (1962, 1969) has provided plausible definitions of large, medium, and small effects and an extensive set of computational aids to the estimation of power for a variety of statistical tests. In the normal test for a difference between two means, for example, a difference of $.25\sigma$ is small, a difference of $.50\sigma$ is medium, and a difference of 1σ is large, according to the proposed definitions. The mean IQ difference between clerical and semiskilled workers is a medium effect. In an ingenious study of research practice, J. Cohen (1962) reviewed all the statistical analyses published in one volume of the *Journal of Abnormal and Social Psychology*, and computed the likelihood of detecting each of the three sizes of effect. The average power was .18 for the detection of small effects, .48 for medium effects, and .83 for large effects. If psychologists typically expect medium effects and select sample size as in the above example, the power of their studies should indeed be about .50.

Cohen's analysis shows that the statistical power of many psychological studies is ridiculously low. This is a self-defeating practice: it makes for frustrated scientists and inefficient research. The investigator who tests a valid hypothesis but fails to obtain significant results cannot help but regard nature as untrustworthy or even hostile. Furthermore, as Overall (1969) has shown, the prevalence of studies deficient in statistical power is not only wasteful but actually pernicious: it results in a large proportion of invalid rejections of the null hypothesis among published results.

Because considerations of statistical power are of particular importance in the design of replication studies, we probed attitudes concerning replication in our questionnaire.

Suppose one of your doctoral students has completed a difficult and time-consuming experiment on 40 animals. He has scored and analyzed a large number of variables. His results are generally inconclusive, but one before-after comparison yields a highly significant $t = 2.70$, which is surprising and could be of major theoretical significance.

Considering the importance of the result, its surprisal value, and the number of analyses that your student has performed, would you recommend that he replicate the study before publishing? If you recommend replication, how many animals would you urge him to run?

Among the psychologists to whom we put these questions there was overwhelming sentiment favoring replication: it was recommended by 66

out of 75 respondents, probably because they suspected that the single significant result was due to chance. The median recommendation was for the doctoral student to run 20 subjects in a replication study. It is instructive to consider the likely consequences of this advice. If the mean and the variance in the second sample are actually identical to those in the first sample, then the resulting value of t will be 1.88. Following the reasoning of Footnote 1, the student's chance of obtaining a significant result in the replication is only slightly above one-half (for $p = .05$, one-tail test). Since we had anticipated that a replication sample of 20 would appear reasonable to our respondents, we added the following question:

Assume that your unhappy student has in fact repeated the initial study with 20 additional animals, and has obtained an insignificant result in the same direction, $t = 1.24$. What would you recommend now? Check one: [the numbers in parentheses refer to the number of respondents who checked each answer]

(*a*) He should pool the results and publish his conclusion as fact. (0)

(*b*) He should report the results as a tentative finding. (26)

(*c*) He should run another group of [median 20] animals. (21)

(*d*) He should try to find an explanation for the difference between the two groups. (30)

Note that regardless of one's confidence in the original finding, its credibility is surely enhanced by the replication. Not only is the experimental effect in the same direction in the two samples but the magnitude of the effect in the replication is fully two-thirds of that in the original study. In view of the sample size (20), which our respondents recommended, the replication was about as successful as one is entitled to expect. The distribution of responses, however, reflects continued skepticism concerning the student's finding following the recommended replication. This unhappy state of affairs is a typical consequence of insufficient statistical power.

In contrast to Responses *b* and *c*, which can be justified on some grounds, the most popular response, Response *d*, is indefensible. We doubt that the same answer would have been obtained if the respondents had realized that the difference between the two studies does not even approach significance. (If the variances of the two samples are equal, t for the difference is .53.) In the absence of a statistical test, our respondents followed the representation hypothesis: as the difference between the two samples was larger than they expected, they viewed it as worthy of explanation. However, the attempt to "find an explanation for the difference between the two groups" is in all probability an exercise in explaining noise.

Altogether our respondents evaluated the replication rather harshly. This follows from the representation hypothesis: if we expect all samples to be very similar to one another, then almost all replications of a valid

hypothesis should be statistically significant. The harshness of the criterion for successful replication is manifest in the responses to the following question:

An investigator has reported a result that you consider implausible. He ran 15 subjects, and reported a significant value, $t = 2.46$. Another investigator has attempted to duplicate his procedure, and he obtained a nonsignificant value of t with the same number of subjects. The direction was the same in both sets of data.

You are reviewing the literature. What is the highest value of t in the second set of data that you would describe as a failure to replicate?

The majority of our respondents regarded $t = 1.70$ as a failure to replicate. If the data of two such studies ($t = 2.46$ and $t = 1.70$) are pooled, the value of t for the combined data is about 3.00 (assuming equal variances). Thus, we are faced with a paradoxical state of affairs, in which the same data that would increase our confidence in the finding when viewed as part of the original study, shake our confidence when viewed as an independent study. This double standard is particularly disturbing since, for many reasons, replications are usually considered as independent studies, and hypotheses are often evaluated by listing confirming and disconfirming reports.

Contrary to a widespread belief, a case can be made that a replication sample should often be larger than the original. The decision to replicate a once obtained finding often expresses a great fondness for that finding and a desire to see it accepted by a skeptical community. Since that community unreasonably demands that the replication be independently significant, or at least that it approach significance, one must run a large sample. To illustrate, if the unfortunate doctoral student whose thesis was discussed earlier assumes the validity of his initial result ($t = 2.70$, $N = 40$), and if he is willing to accept a risk of only .10 of obtaining a t lower than 1.70, he should run approximately 50 animals in his replication study. With a somewhat weaker initial result ($t = 2.20$, $N = 40$), the size of the replication sample required for the same power rises to about 75.

That the effects discussed thus far are not limited to hypotheses about means and variances is demonstrated by the responses to the following question:

You have run a correlational study, scoring 20 variables on 100 subjects. Twenty-seven of the 190 correlation coefficients are significant at the .05 level; and 9 of these are significant beyond the .01 level. The mean absolute level of the significant correlations is .31, and the pattern of results is very reasonable on theoretical grounds. How many of the 27 significant correlations would you expect to be significant again, in an exact replication of the study, with $N = 40$?

With $N = 40$, a correlation of about .31 is required for significance at the .05 level. This is the mean of the significant correlations in the original study. Thus, only about half of the originally significant correlations (i.e., 13 or 14) would remain significant with $N = 40$. In addition, of course, the

correlations in the replication are bound to differ from those in the original study. Hence, by regression effects, the initially significant coefficients are most likely to be reduced. Thus, 8 to 10 repeated significant correlations from the original 27 is probably a generous estimate of what one is entitled to expect. The median estimate of our respondents is 18. This is more than the number of repeated significant correlations that will be found if the correlations are recomputed for 40 subjects randomly selected from the original 100! Apparently, people expect more than a mere duplication of the original statistics in the replication sample; they expect a duplication of the significance of results, with little regard for sample size. This expectation requires a ludicrous extension of the representation hypothesis; even the law of small numbers is incapable of generating such a result.

The expectation that patterns of results are replicable almost in their entirety provides the rationale for a common, though much deplored practice. The investigator who computes all correlations between three indexes of anxiety and three indexes of dependency will often report and interpret with great confidence the single significant correlation obtained. His confidence in the shaky finding stems from his belief that the obtained correlation matrix is highly representative and readily replicable.

In review, we have seen that the believer in the law of small numbers practices science as follows:

1. He gambles his research hypotheses on small samples without realizing that the odds against him are unreasonably high. He overestimates power.

2. He has undue confidence in early trends (e.g., the data of the first few subjects) and in the stability of observed patterns (e.g., the number and identity of significant results). He overestimates significance.

3. In evaluating replications, his or others', he has unreasonably high expectations about the replicability of significant results. He underestimates the breadth of confidence intervals.

4. He rarely attributes a deviation of results from expectations to sampling variability, because he finds a causal "explanation" for any discrepancy. Thus, he has little opportunity to recognize sampling variation in action. His belief in the law of small numbers, therefore, will forever remain intact.

Our questionnaire elicited considerable evidence for the prevalence of the belief in the law of small numbers.[2] Our typical respondent is a believer, regardless of the group to which he belongs. There were practically no differences between the median responses of audiences at a

[2] W. Edwards (1968, 25) has argued that people fail to extract sufficient information or certainty from probabilistic data; he called this failure conservatism. Our respondents can hardly be described as conservative. Rather, in accord with the representation hypothesis, they tend to extract more certainty from the data than the data, in fact, contain.

mathematical psychology meeting and at a general session of the American Psychological Association convention, although we make no claims for the representativeness of either sample. Apparently, acquaintance with formal logic and with probability theory does not extinguish erroneous intuitions. What, then, can be done? Can the belief in the law of small numbers be abolished or at least controlled?

Research experience is unlikely to help much, because sampling variation is all too easily "explained." Corrective experiences are those that provide neither motive nor opportunity for spurious explanation. Thus, a student in a statistics course may draw repeated samples of given size from a population, and learn the effect of sample size on sampling variability from personal observation. We are far from certain, however, that expectations can be corrected in this manner, since related biases, such as the gambler's fallacy, survive considerable contradictory evidence.

Even if the bias cannot be unlearned, students can learn to recognize its existence and take the necessary precautions. Since the teaching of statistics is not short on admonitions, a warning about biased statistical intuitions may not be out of place. The obvious precaution is computation. The believer in the law of small numbers has incorrect intuitions about significance level, power, and confidence intervals. Significance levels are usually computed and reported, but power and confidence limits are not. Perhaps they should be.

Explicit computation of power, relative to some reasonable hypothesis, for instance, J. Cohen's (1962, 1969) small, large, and medium effects, should surely be carried out before any study is done. Such computations will often lead to the realization that there is simply no point in running the study unless, for example, sample size is multiplied by four. We refuse to believe that a serious investigator will knowingly accept a .50 risk of failing to confirm a valid research hypothesis. In addition, computations of power are essential to the interpretation of negative results, that is, failures to reject the null hypothesis. Because readers' intuitive estimates of power are likely to be wrong, the publication of computed values does not appear to be a waste of either readers' time or journal space.

In the early psychological literature, the convention prevailed of reporting, for example, a sample mean as $M \pm PE$, where PE is the probable error (i.e., the 50% confidence interval around the mean). This convention was later abandoned in favor of the hypothesis-testing formulation. A confidence interval, however, provides a useful index of sampling variability, and it is precisely this variability that we tend to underestimate. The emphasis on significance levels tends to obscure a fundamental distinction between the size of an effect and its statistical significance. Regardless of sample size, the size of an effect in one study is a reasonable estimate of the size of the effect in replication. In contrast, the estimated significance level in a replication depends critically on sample size. Unrealistic expectations concerning the replicability of significance levels may be corrected

if the distinction between size and significance is clarified, and if the computed size of observed effects is routinely reported. From this point of view, at least, the acceptance of the hypothesis-testing model has not been an unmixed blessing for psychology.

The true believer in the law of small numbers commits his multitude of sins against the logic of statistical inference in good faith. The representation hypothesis describes a cognitive or perceptual bias, which operates regardless of motivational factors. Thus, while the hasty rejection of the null hypothesis is gratifying, the rejection of a cherished hypothesis is aggravating, yet the true believer is subject to both. His intuitive expectations are governed by a consistent misperception of the world rather than by opportunistic wishful thinking. Given some editorial prodding, he may be willing to regard his statistical intuitions with proper suspicion and replace impression formation by computation whenever possible.