star "makes or breaks" the results. By using large "anonymous" source groups the incentive for any one individual to create false anomalies might be greatly reduced.

STATISTICAL ISSUES AND METHODS*

WHEN WILL WE BEGIN TO REDUCE ALPHA AND BETA ERRORS IN STATISTICAL PSI EXPERIMENTS?

Ulrich Timm (Institut für Grenzgebiete der Psychologie und Psychohygiene, Eichhalde 12, 7800 Freiburg i.Br., West Germany)

In many psi experiments some statistical selection errors are made, after whose correction the initial statistical significance disappears. These are Type I errors, more simply called alpha errors. That does not necessarily mean, however, that in these experiments real psi effects do not exist, since the usual methods, if utilized correctly, are often so ineffective--with regard to the rareness, instability, and inconsistency of psi effects--that they can only seldom lead to statistical significance. This inefficiency of statistical methods creates Type II errors, or beta errors. Therefore, our objective should not only be the reduction of alpha errors and the related decrease of spurious significances but also the reduction of beta errors and the related increase of real significance.

First I give an overview of those alpha errors that I call statistical selection errors. These show, simply stated, the following three qualities (Timm, ZP, 1983, 195-229):

(1) From a set of statistical results a single result is selected and evaluated by some significance test.

(2) The selection is not performed randomly but according to a criterion that is related to the level of the single result in that it directly or indirectly favors positive results.

(3) Despite this success-dependent selection, the significance test is carried out and interpreted in the usual manner without any correction.

*Chaired by Martin U. Johnson.

Following this simple recipe it is almost always possible, even in such investigations whose results are purely random, to find some kind of "significant effects." If one finds, for example, among 20 independent statistical results one single result in excess of the 5% significance limit, then one should correctly ascertain that this corresponds exactly to chance expectation. If one, however, singles out that particular result and declares it as significant, then one will have made an exemplary selection error! In contrast, the correct evaluation would consist in a statistical analysis of the total result. Through such a global significance test every statistical selection error will automatically be avoided. But one can also apply a correction formula to individual selections.

A look at experimental parapsychology immediately shows that it supplies virtually fantastic possibilities to make such selection errors. Already in the evaluation of simple standard experiments containing only one hit variable the following (intra- or interexperimental) selection errors appear with various frequency:

(1) The selection of single temporal sections of an experiment, for example, single "runs," "sessions," "situations," etc.

(2) The selection of single subjects from the total group.

(3) The selection of single significance tests from several tests responding differently to the intraindividual or interindividual score distributions.

(4) The selection of single experiments from the total number of all replications of an experiment.

(5) The selection of single kinds of experiments from the total number of all psi experiments.

However, there seems to be a plausible argument that one would be allowed in parapsychology to test separately the significance of single experimental sections, single subjects, single experiments and so on. One says, namely, that the separate results are not homogeneous because of the great intra- and interexperimental variability of psi performance. Heterogeneous results, one says further, need not be combined since each time one is testing a different hypothesis. Unfortunately, I cannot accept this argumentation: The significance test of a statistical experiment always refers to the null hypothesis; and, in the case of complex experiments, which can be broken down into a number of parts, there usually exists a whole hierarchy of null hypotheses. Then any subordinate null hypothesis is to be interpreted as a special case of a superordinate null hypothesis and can only be rejected if the superordinate null hypothesis has already been rejected. Correspondingly, the subordinate results, in reference to all superordinate

null hypotheses, are to be classified as homogeneous and can only then be separately tested when all of the superordinate results have become significant.

In parapsychology, one can even formulate such a general null hypothesis that it is superordinated to each and every psi experiment. It simply states that psi phenomena do not exist at all. Thus, to evade selection errors, one had to combine all of the psi experiments up to that point and let them undergo a global significance test before one is allowed to interpret them separately. Even if one assumes that, meanwhile, the existence of psi has been established, one must in any case test the total result of every single experiment, since the psi effect is said to vary among experiments and consequently may not necessarily appear in each of them. Only if the total result is significant is one allowed, then, to test the significance of partial results.

The same possibilities of error exist also in the case of differential or correlational psi experiments, which examine differences between various experimental conditions or correlations between psi variables and other variables (e.g., the sheep-goat effect). Here, the same principle of hierarchy is valid: wherever a meaningful superordinate null hypothesis exists, it must be rejected before separate experimental effects, correlations, etc. are allowed to undergo a normal significance test. Therefore, one must also demand the calculation of global significance tests for almost all correlational experiments. In the case of multivariate designs containing many experimental conditions, personality, or psi variables, this can be done through a multiple or canonical correlation in which the psi variables serve as criteria and the other variables as predictors. If one abstains from this, one will find in every larger set of predictor variables some significant correlations with any psi variables; but if one singles them out and interprets them in the usual manner, one makes a selection error and could possibly fall victim to a statistical artifact. If the apparently discovered effect is not replicated in the next experiment, this corresponds to statistical expectation and naturally has nothing to do with the "nonrepeatability" of psi.

One may object to this discussion that sophisticated experiments are carried out in a much more refined manner. Here, in advance, one formulates certain hypotheses which correspond to expected correlations or differences within the results. In the evaluation one limits oneself to these hypotheses. In this case selection errors are said to be excluded and only then possible if one tests post-hoc hypotheses. Unfortunately, this argument is also not completely correct. It is true that one limits the evaluation possibilities through these preformulated hypotheses, which is very recommendable. However, if one has formulated sufficiently enough hypotheses, they still have among these hypotheses enough possibilities for selection. One must, for that reason, here also

Approved For Release 2000/08/11 : CIA-RDP96-00792R000700660002-7

carry out a global significance test for such single hypotheses to which a superordinate null hypothesis can be assigned.

It should be clear that by performing global significance tests many psi experiments must lose their significance. I remember, though, that I also mentioned the interexperimental selection above, to whose avoidance, at the least, all similar psi experiments should be combined and submitted to a global significance test. Through such a "meta-analysis," on the other hand, the significance may increase so that the single experiment loses part of its meaning.

My second theme is the reduction of beta errors in the statistical evaluation of psi experiments. The problem is to increase the statistical efficiency (or power) of the significance tests in such a way that--despite the avoidance of selection errors--minimal psi effects can be statistically detected. I confine myself to two different questions, both of which are of considerable importance to the practice. The first question is: which are the statistically optimal methods for correcting a given selection or for combining single results which shall undergo a global significance test?

Here, it can first be answered that for any selection of a single result there is a simple statistical correction possible that replaces the global significance test. An approximate formula for this purpose requires that one multiplies the p value of the selected result with the number of given results. Naturally, in this manner, the p value will be strongly increased so that the statistical significance will in most cases disappear, as in the case of a global significance test. Nevertheless, this is a universal and very simple method of correcting intra- or interexperimental selection.

Most of the other methods consist in weighted combinations of the single results so as to attain a most efficient global significance test. In the case of standard psi experiments that seems trivial because one needs only to add the different hits, whose sum can be evaluated with a CR just as well as the separate results. However, an analysis of intra- and interindividual distributions of psi scores shows that the simple addition of hits is one of the statistically least efficient methods, even for the aggregation of small experimental units such as individual runs. The reason for this lies in the strong variability of psi scores, which can vary even in a bipolar fashion between psi-hitting and psi-missing so that the hit deviations cancel out each other. Therefore, I have suggested special (nonlinear) transformations weighting the single scores according to their size. Finally, following the method of the likelihood quotient, I came to a measure which is statistically most efficient for strongly varying psi scores and is a linear function of the well-known "run-score variance."

The second question refers to the identification of permissible

forms of selection which one could use to increase the statistical efficiency. For example, the above definition of selection error allows one to exclude any partial results from the global significance test of an experiment if the exclusion ensues according to a criterion that, under the null hypothesis, is independent of the respective results. If one, in this way, discovers certain clues that particular experimental situations, certain subjects, certain variables, etc., could be unsuccessful, one is allowed to eliminate them as is. This can be a great advantage because every nonsignificant partial result reduces the significance of the total result.

In the global statistical evaluation of a multivariate experiment, one should, further, reduce correlated criterion or predictor variables to a smaller number of factors by performing a factor analysis, because the statistical efficiency in the case of correlated variables decreases with the number of variables. Finally, the so-called extreme-group method should be mentioned, according to which one is allowed to eliminate the middle cases of the distribution of a variable when calculating correlations. For example, one could eliminate all the chance-scoring subjects of a correlational study, if enough psi-hitters and psi-missers remain. The correlations between psi variables and other variables could, in that way, become much more significant.

I am afraid my explanations will not lead to a decisive change in the statistical methods of parapsychologists. When I pointed to the problem of statistical selection errors at the 1980 PA Convention in Reykjavik, it also did not have any considerable effect. One must, apparently, turn to the psi skeptics to attain such effects. Probably, selection errors serve the general psychological tendency to synchronize the given empirical data with one's own expectations regarding reality. Therefore, the final demand can only be to answer one's own ways of acting with increased self-criticism, even in such an objective area as mathematical statistics. Otherwise, those cynics will be confirmed who always have contended that, with statistics, one can prove everything.

EVALUATING FREE-RESPONSE RATING DATA

Sybo A. Schouten† and Gert Camfferman (Parapsychology Laboratory, University of Utrecht, Sorbonnelaan 16, 3584CA Utrecht, The Netherlands)

During the recent decades the use of forced-choice methods in experimental research in parapsychology has gradually declined in favor of free-response techniques. A disadvantage of free-response techniques is that they are rather time consuming. The

Approved For Release 2000/08/11 : CIA-RDP96-00792R000700660002-7