Proceedings of the Iowa Academy of Science

Volume 62 | Annual Issue Article 58

1955

A Comparison of Methods for Handling Extreme Values in Experimental Procedure

Virtus W. Suhr
Iowa State College

Let us know how access to this document benefits you

Copyright ©1955 Iowa Academy of Science, Inc.
Follow this and additional works at: https://scholarworks.uni.edu/pias

Recommended Citation
Suhr, Virtus W. (1955) "A Comparison of Methods for Handling Extreme Values in Experimental Procedure," *Proceedings of the Iowa Academy of Science, 62(1),* 473-478.
Available at: https://scholarworks.uni.edu/pias/vol62/iss1/58

This Research is brought to you for free and open access by the Iowa Academy of Science at UNI ScholarWorks. It has been accepted for inclusion in Proceedings of the Iowa Academy of Science by an authorized editor of UNI ScholarWorks. For more information, please contact scholarworks@uni.edu.

A Comparison of Methods for Handling Extreme Values in Experimental Procedure*

By Virtus W. Suhr

Problem

In dealing with certain types of psychological measures, particularly those involving perception and attention, one is sometimes confronted with excessive values which are probably due to lapse of one type or another. Occasionally these extreme values may be numerous enough to spuriously influence the results. This is particularly true when the mean value of a series of scores is taken as a basis for comparing performance of two or more individuals or groups under different experimental conditions.

Several rule-of-thumb methods have been used to secure a valid index for comparison of such measures. Among them are:

1. Use of the median which is not influenced by long responses.
2. Use of the mode as being a typical score.
3. To arbitrarily throw out the unduly long measures as being non-representative or atypical.
4. To average the long measures in with the others assuming that they are representative or typical.
5. To assign an average score to extreme measures.

While the median gives little weight to extreme deviations, it may fail entirely to represent the type. The mode is perhaps in theory the most typical score. However, if an array is very irregular, there is, in strictness, no mode or type at all, or at least the indicated mode has little significance.

To throw out the unduly long measures immediately presents the problem of determining what constitutes a truly extreme score. In other words what should be the criterion against which to judge whether to retain or eliminate a particular score. To average the long measures in with the others does not seem to be satisfactory in that it is likely to unduly influence the mean due to the inclusion of a few extreme values, especially when N is small.

On the other hand it does not seem that extremely long measures should be neglected entirely as they may well be an indicator of some psychological function being measured apart from the immediate scores being sought.

In a laboratory study of driving performance over a three-hour period one of the factors considered was stop-light-response time, that is the amount of time elapsing between presentation, without

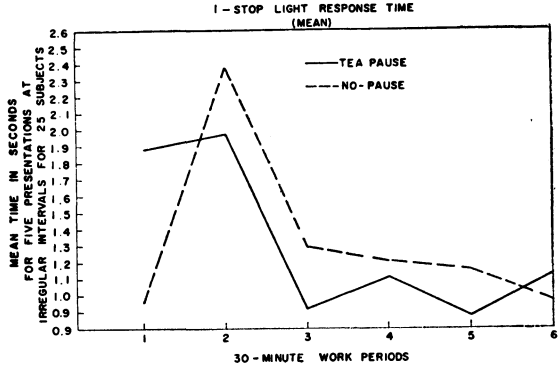
*Thos. J. Lipton, Inc. Project on reducing driving fatigue.

Published by UNI ScholarWorks, 1955

Proceedings of the Iowa Academy of Science, Vol. 62 [1955], No. 1, Art. 58

warning, of a red light in a traffic control device and the depressing of the brake pedal by the driver who had been instructed to stop instantly every time he saw the red light. The traffic lights were located above and to the right of the roadway just as in a regular driving situation. The red light was presented at irregular intervals five times during each half-hour driving period. Two groups of twenty-five subjects each were used in the study. One group drove for three hours straight, the other was given a fifteen-minute rest pause and served refreshments after one and one-half hours of driving.

The mean of the recorded time for the five responses was computed for each subject every half-hour during the driving period. This mean value was taken as the subject's score for the particular half-hour of driving. An over-all mean was computed for each group for each half-hour of the driving period. These means were plotted and the groups were compared graphically as shown in Figure 1.



Inspection of Figure 1 may give the impression of chance fluctuation in the true mean value. However, there is some indication of a consistent difference between the two groups from the second period through the fifth period except as influenced by extremes or outlying scores.

This paper has to do with the comparison of certain correction indexes in an attempt to secure a reasonable method of obtaining the most valid results from data of this nature. Grubbs (3) in discussing sample criteria for testing outlying observations states, "It is clear that rejection of the 'outliers' in a sample will in a great number of cases lead to a different course of action than would have been taken had such observations been retained in the sample. Actually, the rejection of 'outlying' observations may be just as much a practical (or common sense) problem as a statistical one

https://scholarworks.uni.edu/pias/vol62/iss1/58

Suhr: A Comparison of Methods for Handling Extreme Values in Experiment

and sometimes the practical or experimental viewpoint may naturally outweight any statistical contributions." Another investigator, Rider (5), concludes, "In the final analysis it would seem that the question of the rejection or the retention of a discordant observation reduces to a question of common sense. Certainly the judgment of an experienced observer should be allowed considerable influence in reaching a decision. This judgment can undoubtedly be aided by the application of one or more tests based on the theory of probability, but any test which requires an inordinate amount of calculation seems hardly to be worth while, and the testimony of any criterion which is based upon a complicated hypothesis should be accepted with extreme caution." The rationale of these writers was considered in making the described comparisons.

Method and Procedure

Several possibilities were considered for arriving at a satisfactory correction factor for the purposes of the problem at hand. Two types of information seemed to be inherent in the data, viz.;

(a) How long it took the person to respond when he was paying attention—a mean value.

(b) How frequently his attention lapsed—a type of enumerative data.

It was arbitrarily decided that not over 10 per cent of the scores could be eliminated without losing valuable information. Also, the method should not involve an excessive amount of calculation. One index that met these criteria was to use one and one-half times the overall mean for the half-hour period as the critical value. The data were then reanalyzed omitting all scores that exceeded the criterion set up. The adjusted means were then computed for each group for each half-hour of the driving period in the same manner as previously described for the original data and the results were again plotted. Much more consistent curves were obtained.

In order to compare the groups with respect to the frequency of extreme scores, the number of such scores was tabulated for each subject and summed for each group every half-hour. These totals were then plotted and graphed in order to reveal any group differences that might exist. As a further check on the possibility of any real difference between the groups with respect to the number of extreme scores obtained, chi square was computed using the total number of scores obtained each half-hour period. Also, the original means were again adjusted using the Dixon-Massey (2) formula which is based on a contamination model $N(\mu,\lambda^2\sigma^2)$ representing the occurrence of an error from a population with the same mean, but with a greater variance than the remainder of the sample.

Published by UNI ScholarWorks, 1955

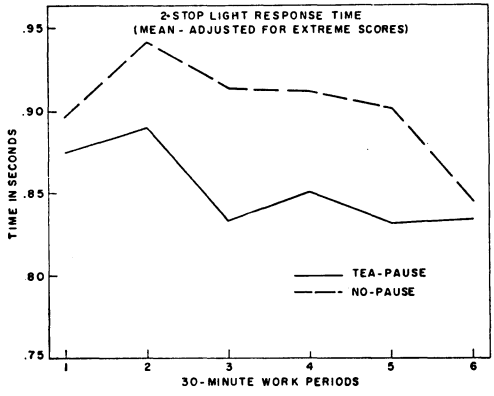
Results

When the means were adjusted for extreme scores, using the criterion of one and one-half times the over-all mean for the period, a consistent difference between the groups was indicated as shown in Figure 2.

Both groups fluctuated considerably with respect to the number of unduly long responses made each half-hour during the entire driving period. This is revealed graphically in Figure 3.

When the Dixon-Massey formula was used for adjusting the means, the dispersion (Figure 4) obtained was somewhat similar to that which resulted from the use of one and one-half times the over-all mean for the period as a correction value. This can be observed by comparing Figure 2 with Figure 4.

The chi-square values for the number of extremely high scores were far below the tabled five per cent level of confidence when using this index of correction.

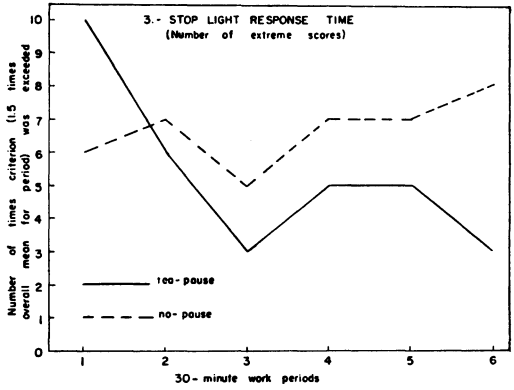


Thus it would appear as expected that adjusting for extreme scores tends to give more consistent results. Differences in the number of such scores does not differentiate the two groups and hence might not be expected to do violence to the results obtained from comparing adjusted mean scores. The two methods give comparable results with no overlapping values.

Summary

In certain types of psychological measures, particularly those involving attention, spurious results may be obtained because of exceedingly long responses due to lapse of perceptual alertness.

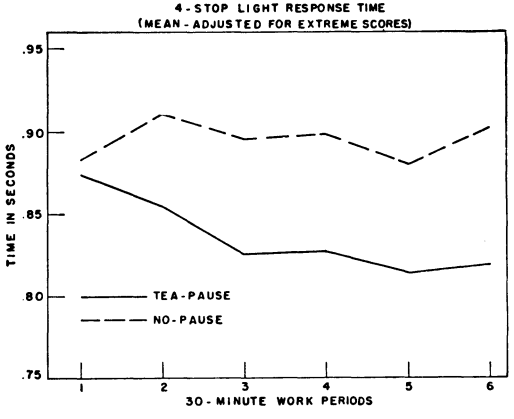
There are several rule-of-thumb methods for dealing with such scores. More precise methods have been used which may be more complicated than the problem demands.



In analyzing a laboratory study of driving performance over a three-hour period, a need arose for a correction index that would give some indication of how long it took a person to respond to a stop-light when not alerted and how frequently his attention might lapse under the experimental conditions imposed.

None of the rule-of-thumb methods seemed quite adequate for this purpose. Several methods devised specifically for adjusting for extreme scores were considered. Finally the criterion of one and one-half times the over-all mean was selected, because it resulted in the elimination of less than 10 per cent of the scores and did not involve any complicated calculations. The scores were adjusted accordingly.

Consistent group differences were brought out when the adjusted scores were plotted and graphed. Similar results were obtained when the Dixon-Massey formula was used to adjust the scores for



Proceedings of the Iowa Academy of Science, Vol. 62 [1955], No. 1, Art. 58

extreme values. A chi-square test of differences in the number of extreme values yielded no tabled significant difference.

The method used is not proposed as a solution to all problems, but it seemed to have merit as an evaluating index in the analysis of our data. At least it seemed to remove some of the variance found with a minimum amount of calculation and agreed closely with the results obtained by application of the Dixon-Massey method which is recommended for general application.

References

1. Dixon, W. J., Analysis of extreme values. The Annals of Mathematical Statistics, 21(no. 4):488-506, December 1950.
2. Dixon, Wilfred J. and Massey, Frank J. Jr., Introduction to Statistical Analysis. New York, McGraw-Hill Book Company, Inc., 1951.
3. Grubbs, Frank E., Sample criteria for testing outlying observations. The Annals of Mathematical Statistics, 21(no. 1):27-58, March 1950.
4. Lewis, Don and Burke, C. J., The use and misuse of the chi-square test. Psychological Bulletin, 46(no. 6):433-489, November 1949.
5. Rider, P. R., Criteria for rejection of observations. St. Louis, Washington Univ. Studies—New Series, Science and Technology—No. 8, 1933. (Original not examined—cited in 3).
6. Wert, James E., Neidt, Charles O., and Ahmann, J. Stanley, Statistical Methods in Educational and Psychological Research. New York, Appleton-Century-Crofts, Inc., 1954.
7. Whipple, Guy Montrose, Manual of Mental and Physical Tests. Baltimore, Warwick and York, Inc., 1914.

Driving Research Laboratory
Iowa State College
Ames, Iowa

https://scholarworks.uni.edu/pias/vol62/iss1/58