ANALYSIS OF EXTREME VALUES

By W. J. Dixon[1]

University of Oregon

1. Introduction. It is well recognized by those who collect or analyze data that values occur in a sample of n observations which are so far removed from the remaining values that the analyst is not willing to believe that these values have come from the same population. Many times values occur which are "dubious" in the eyes of the analyst and he feels that he should make a decision as to whether to accept or reject these values as part of his sample. On the other hand he may not be looking for an error, but may wish to recognize a situation when an occasional observation occurs which is from a different population. He may wish to discover whether a significant analysis of variance indicates an extreme value significantly different from the remainder. Also, of course, the extreme value may differ significantly without causing a significant analysis of variance and he may wish to discover this. It is reasonable to suppose that a criterion for rejecting observations would be useful here also. The choice of a suitable criterion for rejecting observations introduces a number of questions.

1. Should any observations be removed if we wish a representative sample including whatever contamination arises naturally? In other words, it may be desirable to describe the population including *all* observations, for only in that way do we describe what is actually happening.

2. If the analyst wishes to sample the population unaffected by contamination he must either remove the contaminating items or employ statistical procedures which reduce to a minimum the effect of the contamination on the estimates of the population. That is, he may wish to describe only 95% of his population if the description is altered radically by the remaining 5% of the observations. He may have external reasons which are good and sufficient for wishing to describe only 95% of his observations. Suppose he wishes to use the sample for a statistical inference; the inclusion of all the data may sufficiently violate the assumptions underlying the inference to exclude the possibility of making a valid inference.

This paper will concern itself only with those problems which arise from Question 2.

If we wish to follow some procedure which attempts to remove contamination we must consider the performance of any proposed criterion with respect to the proportion of contamination the criterion will discover and, of course, the proportion of the "good" observations which are removed by the use of the criterion. But, perhaps more important, we must consider what sort of bias will result when the standard statistical procedures are applied to samples of observations which have been processed in this manner.

[1] This paper was prepared under a contract with the Office of Naval Research.