



Taylor & Francis
Taylor & Francis Group



Combining Overlapping Information

Author(s): Richard Zeckhauser

Source: *Journal of the American Statistical Association*, Vol. 66, No. 333 (Mar., 1971), pp. 91-92

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: <https://www.jstor.org/stable/2284854>

Accessed: 28-05-2020 19:07 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



JSTOR

Taylor & Francis, Ltd., American Statistical Association are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association*

Combining Overlapping Information

RICHARD ZECKHAUSER*

It is frequently necessary to combine the information provided by summary statistics for different observation sets, particularly in the context of group decision processes. Problems arise when there are unidentified observations that overlap, that is, belong to more than one set. This article discusses the handling of these problems in relation to a Normal example when the exact extent of overlap is known.

1. INTRODUCTION

In decision situations it is frequently necessary to combine the information derived from different sets of observations. Difficulties arise when there is some overlap in these sets, but the data cannot be sifted to identify the overlapping observations.

An individual might, for instance, possess a summary statistic for each of two populations (or for two samples drawn from a single population) which have some elements in common, and he might wish to estimate the summary statistic for the two populations combined. The populations might be the residents of overlapping geographical regions or the members of different groups with some common members.

In group or organizational decision contexts, it is frequently desirable to combine the information derived from sets of observations taken by different individuals. Consider the oral examiners who must decide on a grade for a Ph.D. candidate. None of them is fully competent to evaluate all of the answers given to the questions asked by the others. Thus, the examiners have some distinct and some common information. It would be embarrassing as well as most difficult to determine which observations were taken in common; thus, the individuals cannot distill separate information about those observations that overlap.

Information-combining methods are especially needed because experiments "indicate that when *opinions are involved*, face-to-face discussion may, more often than not, result in a group opinion which is less accurate than simply the average of the individual opinions without discussion." (This result has led to the development of the Delphi group decision procedures, whose basic characteristics are: (a) Anonymity, (b) Iteration with controlled feedback, (c) Statistical group response.)¹

We assume that the data are given as summary statistics for the observation sets of different individuals.

Individual i supplies us with a vector of summary statistics, z_i , which is sufficient for the n_i observations that he records by himself. However, if the information from the different individuals' observation sets is to be pooled, summary statistics will be insufficient unless they enable us to identify the observations that overlap. (Details such as the time and place that observations are taken as well as the exact value of each one may be of use in furthering identification.)

The purpose of combining the information supplied by different individuals is to determine a likelihood function for θ , a parameter whose distribution we wish to assess. Here likelihood function means any function of θ proportional in θ to the probability that the observed data are recorded conditional upon θ . That is,

$$\ell(\theta|\text{data}) \propto f(\text{data}|\theta).$$

2. NORMAL EXAMPLE WITH EXACT EXTENT OF OVERLAP KNOWN

As a simple and concrete example of an information-combining problem, consider the case where two investigators have taken samples of size n_1 and n_2 from a common normal distribution with known unit variance, but with unknown mean θ . They recorded the same observations in m instances, though which specific observations fell into their overlapping sample is not known. The observed means for their two samples were z_1 and z_2 . The investigators wish to derive the likelihood function for the true mean, θ , of the normal distribution conditional upon the means recorded for their two samples, in cognizance of the existence of overlap.

It turns out, quite conveniently, that the likelihood function associated with the described sample is the same as one arising from a single sample of size N with mean Z taken from a normal population with mean θ and variance 1 where

$$N = n_1 + n_2 - m - \eta, \quad (2.1a)$$

$$Z = [z_1(n_1 - m) + z_2(n_2 - m)]/[n_1 + n_2 - 2m], \quad (2.1b)$$

and

$$\eta = 1 / \left(\frac{1}{n_1 - m} + \frac{1}{n_2 - m} + \frac{1}{m} \right). \quad (2.1c)$$

The number of distinct observations in the combined sample is equal to $n_1 + n_2 - m$. Thus N , which in some sense is the size of the effective sample, is equal to the

* Richard Zeckhauser is associate professor of political economy, Harvard University. The author is grateful to Professors John Pratt, Howard Raiffa and Robert Wilson, a referee and an associate editor for their useful comments. Professor Pratt provided a streamlined version of the author's derivation for the Normal example.

¹ See [1]. See also RAND publications RM-5888, RM-5957, P-2982, P-3721.

number of distinct observations minus the quantity η which can be thought of as the compensating decrease in effective sample size due to the fact that overlapping observations are not identified. It is interesting to note that the equivalent mean, Z , for the effective combined sample is a weighted average of the means of the individual samples, with weights proportional to the number of observations unique to each investigator.

For example, if $n_1=30$, $z_1=3.93$, $n_2=40$, $z_2=4.25$, and there is an overlap of $m=10$ unidentified observations, then the likelihood function will be the same as one derived from a sample of size 54.545, with a mean of 4.122. The relative weights assigned to the two means, z_1 and z_2 , are .4 and .6. The decrease in effective sample size, η , is 5.455.

One way to derive the result given in (2.1) is to observe that $n_1 z_1$ and $n_2 z_2$ are sums with m observations in common and, hence, are jointly normal with means $n_1 \theta$, $n_2 \theta$, variances n_1 , n_2 , and covariance m . It follows that

$$\ell(\theta | z_1, z_2) \propto f(z_1, z_2 | \theta) \propto e^{-1/2 Q}, \quad (2.2)$$

where

$$Q = \frac{n_1 n_2}{n_1 n_2 - m^2} [n_1 (z_1 - \theta)^2 - 2m (z_1 - \theta)(z_2 - \theta) + n_2 (z_2 - \theta)^2].$$

Regarding Q as a quadratic form in θ and completing the

square yields

$$\ell(\theta | z_1, z_2) \propto e^{-1/2 N (Z - \theta)^2} \quad (2.3)$$

where N and Z have the forms shown in (2.1).

3. QUALIFICATIONS AND CONCLUSIONS

Real-life decision problems will be more complex than the tractable example discussed in this article. The most frequent case in which information must be combined occurs when individuals, and there may be more than two, who have different opinions on a subject wish to combine their information (probably possessed in some non-quantified form) to arrive at some consensus. They will encounter some familiar, persistent problems such as taking account of differences in individuals' reliabilities, overcoming systematic biases in individual assessments, and separating out common prior information. They will also encounter many difficulties not often discussed in the statistical literature. If we are ever to learn how two heads should best be brought together, these difficulties too will have to be investigated.

REFERENCE

- [1] Dalkey, Norman, *Predicting the Future*, P-3946, RAND Corporation, 1968, 7-8.