PROJECT AIR FORCE

THE ARTS
CHILD POLICY
CIVIL JUSTICE
EDUCATION
ENERGY AND ENVIRONMENT
HEALTH AND HEALTH CARE
INTERNATIONAL AFFAIRS
NATIONAL SECURITY
POPULATION AND AGING
PUBLIC SAFETY
SCIENCE AND TECHNOLOGY
SUBSTANCE ABUSE
TERRORISM AND HOMELAND SECURITY
TRANSPORTATION AND INFRASTRUCTURE
WORKFORCE AND WORKPLACE

This PDF document was made available from www.rand.org as a public service of the RAND Corporation.

Jump down to document

The RAND Corporation is a nonprofit research organization providing objective analysis and effective solutions that address the challenges facing the public and private sectors around the world.

Support RAND

Purchase this document
Browse Books & Publications
Make a charitable contribution

For More Information

Visit RAND at www.rand.org
Explore RAND Project AIR FORCE
View document details

Limited Electronic Distribution Rights

This document and trademark(s) contained herein are protected by law as indicated in a notice appearing later in this work. This electronic representation of RAND intellectual property is provided for non-commercial use only. Unauthorized posting of RAND PDFs to a non-RAND Web site is prohibited. RAND PDFs are protected under copyright law. Permission is required from RAND to reproduce, or reuse in another form, any of our research documents for commercial use. For information on reprint and linking permissions, please see RAND Permissions.

This product is part of the RAND Corporation technical report series. Reports may include research findings on a specific topic that is limited in scope; present discussions of the methodology employed in research; provide literature reviews, survey instruments, modeling exercises, guidelines for practitioners and research professionals, and supporting documentation; or deliver preliminary findings. All RAND reports undergo rigorous peer review to ensure that they meet high standards for research quality and objectivity.

TECHNICAL REPORT

Using Linear Programming to Design Samples for a Complex Survey

James H. Bigelow

Prepared for the United States Air Force

Approved for public release; distribution unlimited

RAND PROJECT AIR FORCE

The research described in this report was sponsored by the United States Air Force under Contract FA7014-06-C-0001. Further information may be obtained from the Strategic Planning Division, Directorate of Plans, Hq USAF.

Library of Congress Cataloging-in-Publication Data

Bigelow, James H.
Using linear programming to design samples for a complex survey / James H. Bigelow.
p. cm.
Includes bibliographical references.
ISBN 978-0-8330-4163-0 (pbk. : alk. paper)
1. United States. Air Force—Public opinion. 2. Linear programming. 3. Social surveys—United States—Case studies. I. Title.

UG633.B54 2007
358.4'1330723—dc22

2007031268

The RAND Corporation is a nonprofit research organization providing objective analysis and effective solutions that address the challenges facing the public and private sectors around the world. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

RAND® is a registered trademark.

© Copyright 2007 RAND Corporation

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from RAND.

Published 2007 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
RAND URL: http://www.rand.org
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Summary

This report describes a method we developed to design and select the sample of Air Force personnel who would be asked to participate in a survey on cultural attitudes. The survey is long and complex and has many competing goals, so the design problem has no simple answer. In the simplest surveys, one seeks only to estimate a population parameter—say, the proportion of the population that prefers chocolate to vanilla—from the responses of the people in the sample. But, in this case, the design considered the following:

- We intended to build multiple three- and four-way crosstabs from the responses (e.g., by grade and AFSC family,[1] and by grade and gender), so we needed to select a sample from which we could estimate differences in population parameters in different three- or four-way crosses with the desired precision.
- Because the survey was very long, we split the sample into three sections and asked the people assigned to each section to respond to only part of the survey. Thus, we needed to trade off sample size (and therefore precision) in one section versus the others.
- Finally, our survey (which we refer to as the CULTURE survey) was fielded shortly after another survey (the HEALTH survey), and because we feared that survey fatigue would reduce the response rate among people in both samples, we wanted to design a sample for our survey that overlapped as little as reasonably possible with the other survey's sample.

Our method partitioned the population into cells. By the definition of a partition, each member of the population belonged to a unique cell. We formed the cells by taking the eight characteristics needed to define all the three- and four-way crosstabs mentioned above and using them to define one gigantic eight-way cross classification. There were over a million cells. More than 98 percent of them were empty, and almost half the nonempty cells contained only one person. For example, there just are not very many company-grade female, Hispanic, Roman Catholic pilots in the Air National Guard who are assigned to the Air Combat Command. (Our method ignores the empty cells, since they contain no personnel to select for our sample.)

We recovered a particular three- or four-way crosstab from the cells (i.e., the eight-way classification) by summing over characteristics not used to define that crosstab. Aggregate data

[1] The Air Force Specialty Code (AFSC) is a code for the skills an individual possesses.

obtained in this way are called *marginal* data, and we will refer to each three- or four-way cross as a *marginal cross*.

Our method used linear programming, a well-known procedure for optimizing (maximizing or minimizing) a linear function subject to linear constraints. The primary variables in our linear program were the expected number of people drawn from each cell for assignment to each section of the survey. (For large cells, there was no practical difference between the number drawn and the expected number. For a cell containing only one person, the expected number was the probability that the person was in the sample.) The constraints ensured that

- the number of people drawn from a cell could not exceed the cell population
- the sample taken from each marginal cross was at least as large as a specified quantity
- the sample taken from each cell was at least as large as another specified quantity. This quantity was selected to ensure that an adequate number of people from the cell were in the sample for every marginal cross to which the cell contributed.

The last two constraint types ensured that estimates of population parameters would have, as far as possible, the desired precision.

Merely because our method used linear programming—which solves an optimization problem—did not mean that we regarded the sample it designed to be "optimal" in a real-world sense. The CULTURE and HEALTH surveys each had its own desired precision, as did each of the sections within the CULTURE survey, and we could only improve the precision of one survey/section at the expense of another.[2] So we embedded a number of "design levers" in the linear program. By manipulating them, we caused the method to generate sample designs that gave different priorities to the various surveys/sections. Flesh-and-blood humans examined the various designs and, ultimately, picked one.

To aid in this process, we defined a number of summary measures of design quality. An obvious one was total sample size. Others were counts of "problem crosses," i.e., marginal crosses for which the sample did not satisfy the last two constraint types and hence would not yield the desired precision. By varying the threshold at which we considered a cross to be a problem, we could construct a whole family of measures.

It seemed to the survey team that the HEALTH survey had rather little precision unless it received a very high priority, but this caused the CULTURE survey to lose too much precision. This observation persuaded the survey team to investigate the effect of allowing the two samples to overlap. Indeed, giving the HEALTH survey a high priority causes the CULTURE survey to lose much less precision in the samples with overlap than in those without.

The last step in the process was to select an actual sample. The linear program solution that we ultimately chose gave us the expected number of people to draw from each cell for each section. We developed a variant of systematic sampling to choose the actual individuals to include in the sample. In our method, each member of the population had some chance of

[2] The CULTURE survey was partitioned into three sections, so assigning a person to a section also assigned that person to the CULTURE survey; anyone assigned to the CULTURE survey had to be in one of the three sections. The HEALTH survey was not divided into sections, so we do not use the term "section" when referring to the HEALTH survey. A person assigned to the overlap was assigned to both the HEALTH survey and a section within the CULTURE survey.

being assigned to any survey/section. The sample our method selected matched the linear programming solution quite closely. Both the overall sample and the sample taken from each large cell had sizes close to the expected sizes calculated by the linear program.

There is no guarantee, however, that the precision of estimates we calculate when we analyze the survey responses will equal the precision we sought when we designed the sample. The actual precision will depend on factors that were unknown when we designed the sample (for example, the actual response rate). Moreover, we could only design the sample to ensure adequate precision of estimates we anticipated we were going to make. Once we begin analyzing the survey responses, we may discover that some quantities we did not anticipate estimating are much more interesting.

In short, we will inevitably judge the adequacy of the sample using standards that we do not fully know at the time we design the sample. We can only guess what those standards will be, design the sample to our guess, and let the chips fall where they may.