

Using Statistics to Find the Dollars

How an understanding of "variables" can help you unlock new possibilities in your donor database.

By Peter B. Wylie

If you took a statistics course in college or graduate school, all you may remember about it are the expensive textbooks and the arcane equations with no apparent application. Even so, if you're a fundraising professional—especially one who works in annual giving or prospect research—you probably realize that if you knew more about statistics, you could make much better use of your donor databases.

What do fundraising professionals really need to know about statistics? Essentially, they need to know a little about the concept of sampling; they need to know a little about variables; and they need to know about the relationships among variables. That is where they'll get the most bang for their buck.

What follows is an introduction to the concept of variables—those things that people (donors and prospective donors in this case) "vary" on. Your donor database undoubtedly has a field called "total amount." This field lists the total dollars each individual has given the organization since that person's record has been in the database. As you well know, total amount varies widely from one record to another. Many people (far more than you'd like) have given absolutely nothing. A good number are likely to have given up to a total of \$100. And a few (very few) have given huge amounts that, for some institutions, can be in the million-dollar-plus range.

What are some of the variables in your database that might be related to these differences in giving? Here are just a few:

- **Business phone.** People with a business phone listed in the database are often more likely to have given than people without a business phone listed.
- **E-mail address.** People with an e-mail address listed in the database may be more likely to have given than people without an e-mail address listed.
- **Marital status.** In many databases, this field tends to be about 50% to 60% populated (which is to say that somewhere between 40% and 50% of the records have no marital code listed). What I generally find is

—— Page 1 of 6 ——

Copyright © 2003, CASE. This article may not be reprinted, reproduced, or retransmitted in whole or in part without the express written consent of the author.

Reprinted here by permission given to The Grantsmanship Center.

<http://www.tgci.com> (800) 421-9512 [Join Our Mailing List](#)

that people who have a marital code listed (regardless of what it is) have given more money and more often than those with no code listed at all. Beyond that, I usually find that people who are listed as "married" or "widowed" (as opposed to "single" or "divorced/separated" or no code at all) give the largest amounts and give the most frequently.

In fact, there are probably lots of variables in your donor database—we've only scratched the surface here—and some of those variables are related to the very important variable of giving.

Although statisticians have never fully agreed on all the different types of variables, there are two primary types: categorical and quantitative.

- **Categorical variables.** These are variables for which it makes no sense to say that one category of data is "more" or "less" than another. Take the field "PREFIX" in a database. The "PREFIX" field has categories such as "Mr.," "Ms.," "Mrs.," "Dr.," and so on. We can't say that "Mr." is more than "Ms." or that "Dr." is less than "Mrs." All we can say is that people vary in terms of their prefixes. Other examples of categorical variables are things like hair color or religion.
- **Quantitative variables.** For these variables, it does make sense to say that one category is more or less than another. Age is a quantitative variable because it is measured in years; age 50 is more than age 49. Weight is a quantitative variable, because 175 pounds is more than 174 pounds. Variables measured in dollars are also typically quantitative variables—that is, \$1,000 is clearly more than \$999.

One good way to decide whether a variable is categorical or quantitative is to ask: "Would it make sense to compute an average (the technical term is mean) for this variable?" If the answer is yes, the variable is almost certainly quantitative; otherwise it's almost certainly categorical.

Let's look at two of the variables in a typical development sample and decide which are categorical and which are quantitative.

- **ID:** This is a unique number that the institution uses to identify each record in its database. On the surface, it looks like a quantitative variable because all the entries are numbers. But if we ask our question—"Would it make sense to compute an average for this variable?"—the answer is obviously no, just as it wouldn't make sense to compute an average for the numbers on the jerseys of basketball

players. These numbers are no more than a convenient way to identify people; they don't imply any sense of "more than" or "less than." So ID is a categorical variable.

- **TOTAL_AMT:** This field lists the total dollars each individual has given the institution since that person's record has been in the database. Since it clearly makes sense to compute an average for this one, we can be pretty sure that TOTAL_AMT is a quantitative variable.

Hybrid Variables

As far as I know, 'hybrid variable' is not a term you'll find in any statistics textbook. It's a term I made up to deal with a problem I often encounter. For example, consider the variable NUM_CHILDREN, which tells us the number of children a person in the database has.

Since it would certainly be reasonable to compute an average for this one, is there any reason we shouldn't call it a quantitative variable? Maybe, but what do we do about the blanks? Do we assume that any person with no entry for NUM_CHILDREN has no children and therefore enter zero for that record?

I don't like that solution because we simply don't know the facts. All we know is that there is nothing listed for the record. So I'm much more inclined to code the blanks as "not listed" or "DK" (don't know). And there's the rub. If I do that, then we have a variable that has both "numeric" and "alpha" data from the standpoint of how the statistical software deals with the variable. That is, if we want to compute an average for NUM_CHILDREN, we can only do it for those records that have a number coded; we can't do it for the records that have a code of "not listed" or "DK."

So we end up with a variable that is both quantitative and categorical—a hybrid variable. These don't get talked about in statistics textbooks, but they're definitely a fact of life in the kind of data analysis work that fundraisers generally do.

Summarizing Variables

We've covered two different types of variables and we've talked about my idea of a hybrid variable. Now we can talk about some ways to summarize these variables. But first let's discuss why summarizing variables is a good thing to do.

Consider the variable MARITAL_STATUS, which indicates whether the person is listed as "divorced," "married," "single," "surviving spouse," "unmarried" or "widowed." Here's how the first nine records might look:

MARITAL_STATUS

married
married
married
married
married
married
unmarried
unmarried
married

Imagine looking at this variable as a field in Excel. If we scroll down through the thousands of records in the file, we'll see entries for "divorced," "married," "single," and so on slide by. While it's useful to glance at all this raw data to get a sense of how it's stored, it's very hard get a handle on all those entries.

But what if we constructed a table like this one that shows the frequencies and percentages of each marital code in our development sample?

MARITAL STATUS	COUNT	%
divorced	8	0.16
married	2935	58.87
single	2	0.04
surviving spouse	1	0.02
unmarried	2031	40.73
widowed	9	0.18
TOTAL	4986	100.00

Suddenly, the mass of data becomes a little easier to grasp. At a glance, we can learn a number of facts about our database:

- Almost 60% of the records list "married."
- About 40% list "unmarried."
- The other categories (as a total) constitute less than one-half of 1% of our database.

Those are facts. Nobody can dispute them unless they find an error in our computations. But a summary like this does more than reveal facts. It stimulates interesting questions. If we're looking for predictors of giving, we would look at this table and ask: "How do the 'marrieds' differ from the 'unmarrieds' in terms of giving? Do the former give more than the latter, or vice versa? Does age affect the difference in giving between these two groups? For example, do young 'marrieds' give less than young 'unmarrieds'—perhaps because they have less disposable income? Does that pattern change as people get older and their kids leave the nest?"

So summarizing a variable allows us to:

- see the variable more as a whole than as a mass of data
- uncover important facts about the variable
- raise important questions about the relationship of the variable to other variables.

What's the best way to summarize variables? As far as I'm concerned, if you can make a table or chart that shows the percentage distribution of a variable and a manager can understand it, then you know how to summarize a variable. The manager who's going to look over your results wants to understand the point of what you're trying to say, not the details of how you developed your point. So let's say you construct a table or chart that shows a percentage distribution of a variable in your database and you give it to the manager. If he or she looks intrigued (rather than confused or bored), you've probably succeeded.

Most managers have a much easier time with charts than they do with tables. The old adage that a picture is worth a thousand words (or, in this case, a thousand numbers) seems to apply here. If you use charts to "draw a picture of your data, "you'll probably make your point—especially when your goal is to convey information to someone who has neither the time nor the interest (nor, perhaps, the aptitude) to deal with a lot of numerical detail.

(If you're going to be spending much time analyzing data, however, you'll probably have to familiarize yourself with tables. Tables are certainly harder to read than charts. But they have at least two distinct advantages over charts: they convey more information in less space and they make it easier for others to check and replicate your work.)

Obviously, there's a lot more involved in using statistics as a data-mining tool than we have covered here. But the fact is that a simple understanding of variables, along with a grasp of sampling, can go a long way in facilitating better, more accurate decisions about your donors and prospective donors. For example, some basic statistics work could help you accomplish things like these:

- If you're picking out promising donors to assess for major gift potential, you can determine which of the thousands of names in your database to submit for a wealth capacity screening.
- If your mailing list has 50,000 names but your budget is only big enough for a mailing to 30,000, you can figure out what portion of your "lybunts" (the people who gave last year but not this year) should get mailings on this round.
- You can identify the major donors to whom your executive director and gift officers should be devoting the bulk of their travel time.
- You can decide who among your 60-and-over widowed supporters are most likely to make a bequest to the organization, and target your pitches accordingly.

With a working knowledge of basic concepts like variables, you'll be well on your way to accomplishing these goals more easily and more effectively.

Peter B. Wylie, Ed.D., is an individual psychologist and data analyst who teaches development professionals how to mine their own databases to find predictors of giving. He can be reached by e-mail at Pbradwylie@aol.com or by phone at (202) 332-7571. This article is adapted from his new guidebook, Data Mining For Fund Raisers, published by the Council for Advancement and Support of Education (CASE). Copyright © 2003, CASE. Reprinted by permission. To order Data Mining For Fund Raisers, or for more information about this and other CASE publications, phone CASE at (202) 328-2273 or visit www.case.org.