STATS

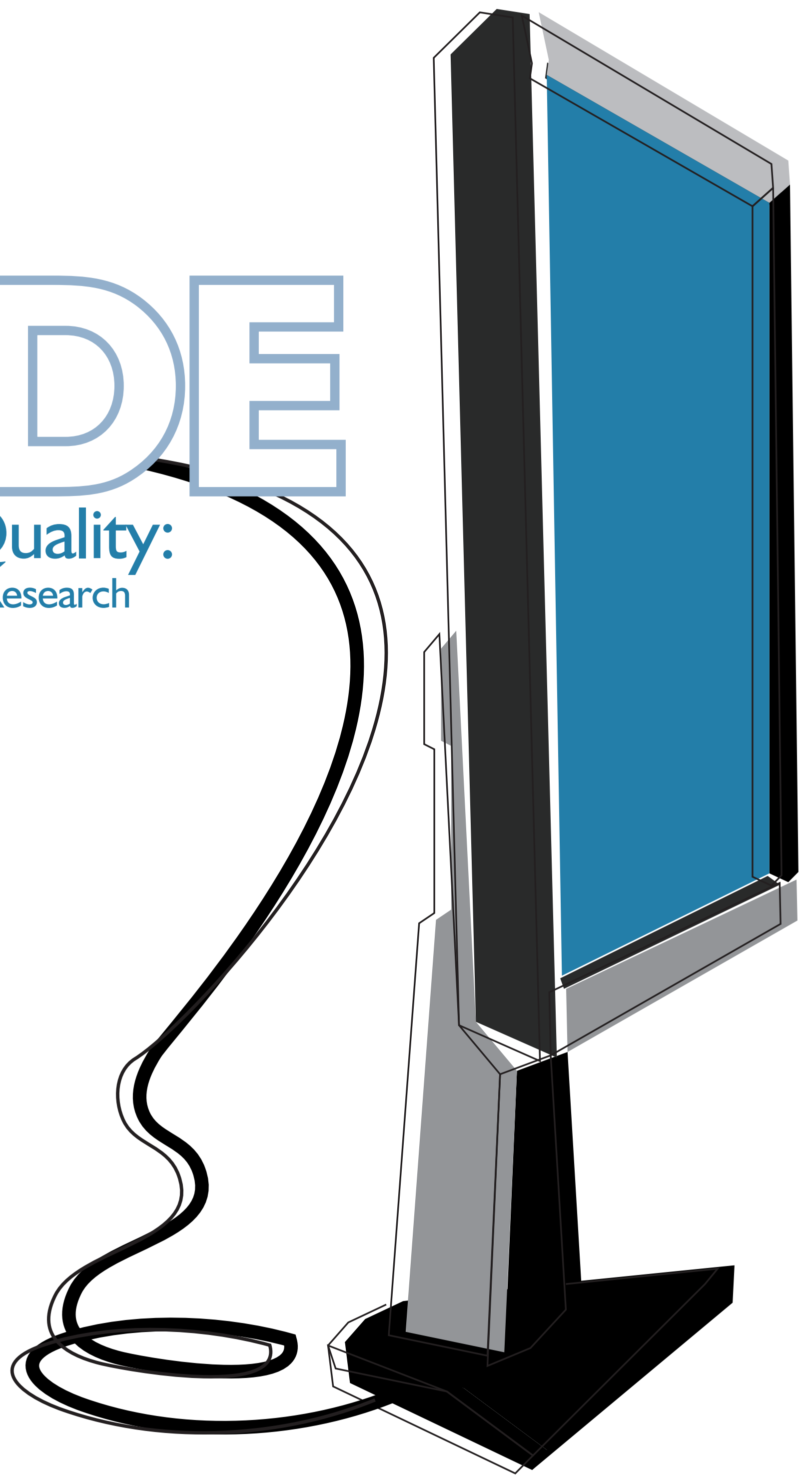
THE MAGAZINE FOR STUDENTS OF STATISTICS : : ISSUE 51

Vince Lampone
gives a quick guide to getting the best online data

Bruce Trumbo
Accept, Reject, ... or Get More Data? An Intuitive View of Sequential Tests

A QUICK GUIDE to Online Data Quality:

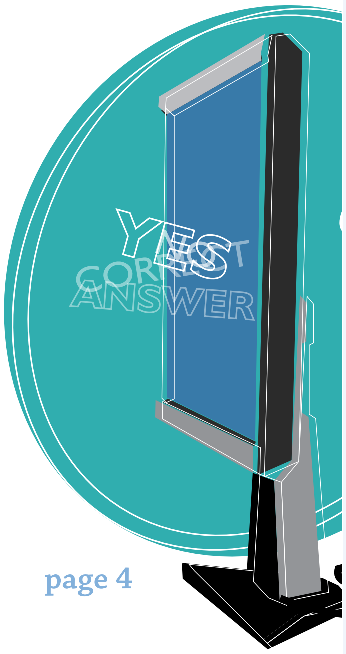
Ensuring High-Quality Data from Panel Research



STATS contents

features

SPECIAL TOPICS



page 4

page 18

A Quick Guide to Online Data Quality: Ensuring High-Quality Data from Panel Research

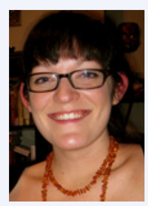
VINCENT LAMPONE is research manager for National Public Radio.

Accept, Reject, … or Get More Data? An Intuitive View of Sequential Tests

BRUCE TRUMBO is a professor of statistics and mathematics at California State University, East Bay. He is a Fellow of the American Statistical Association and a holder of the ASA Founder's Award.

Exit Polling in DC: The 2008 Presidential Election

KATE SHOEMAKER earned her BS in mathematics with a concentration in statistics at Salisbury University before completing her graduate degree in survey design and data analysis at The George Washington University. She works at the Bureau of Economic Analysis as a mathematical statistician.



TIFFANY THOMPSON enrolled in the survey design and data analysis graduate certificate program at The George Washington University after earning her MS in applied demography from Florida State. She is working as a statistician/ demographer at the U.S. Census Bureau.

STATS: *The Magazine for Students of Statistics* (ISSN 1053-8607) is published by the American Statistical Association, 732 North Washington Street, Alexandria, VA 22314-1943 USA; (703) 684-1221; *www.amstat.org*.

Requests for membership information, advertising rates and deadlines, subscriptions, and general correspondence should be addressed to the ASA office.

Dear Readers,

This is the last issue of *STATS*, as well as the second (and last) issue Jill Lacey and I will have had the opportunity to guest edit. We have had fun and will miss seeing *STATS* before you do.

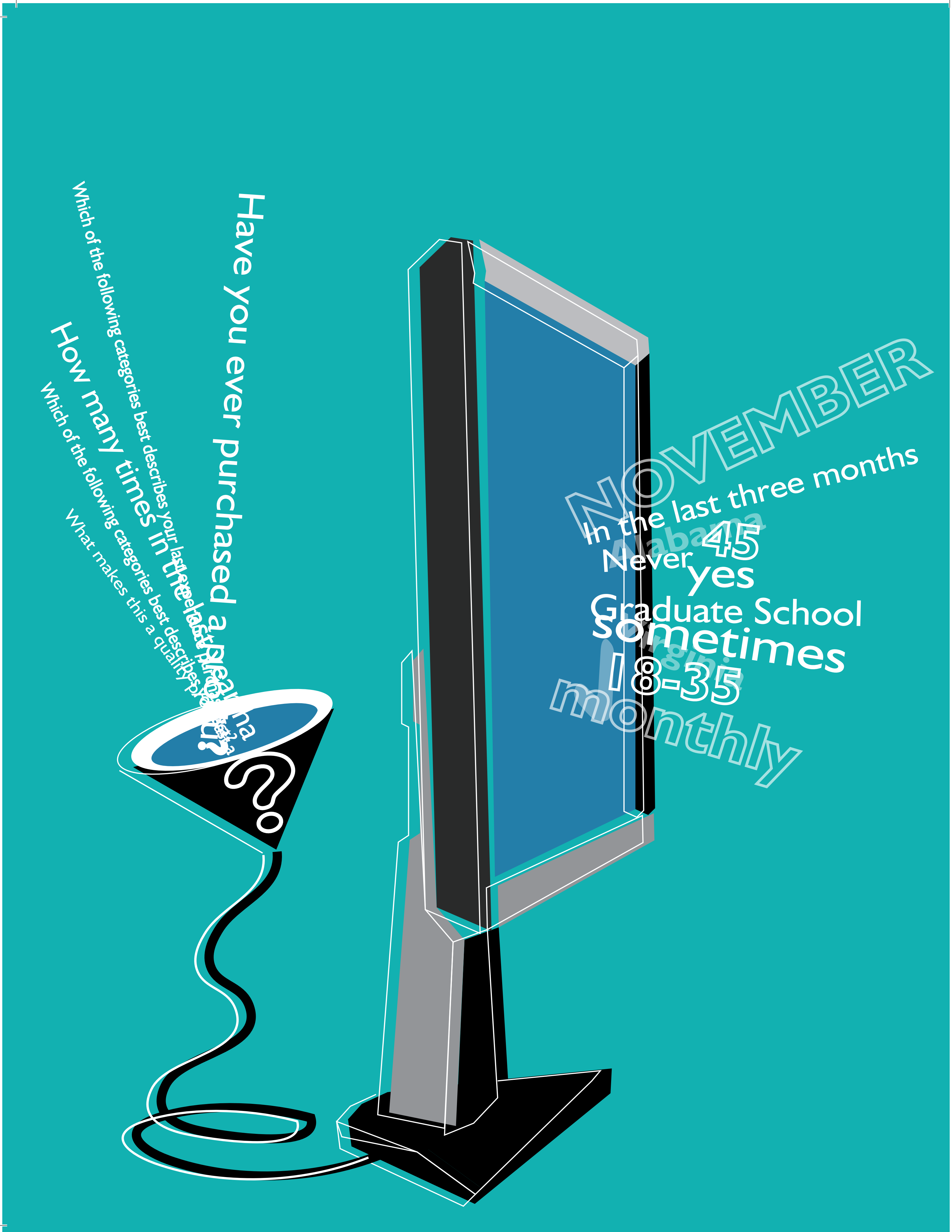
There are three articles included in this issue. We begin with a long piece by Vince Lampone about the use of online data sources not obtained by a conventional survey or census data-collection paradigm. You may find Lampone's work epistemologically challenging, but we recommend it to you in any case. The need to use broader inference supports for information is something we are grappling with as a profession. This article invites you to see how it feels and may urge you to move us all into the "what's next?" world that is coming—or is already here.

The second piece is by Bruce Trumbo and two of his students. It continues the regular *STATS* tradition of looking at computing tools and applications. The article is instructive and needs no advertising, but we do want to call your attention to Trumbo's note, which sums up what *STATS* has meant for many us over the years.

In the final article, we return to the 2008 presidential election and recount an exit poll case study, done by two GWU graduate students in the District of Columbia. There are many points you will find of interest in their work, especially if you intend to use exit poll data or mount an exit poll yourself. Not only is the article a good example of how to collect and handle statistical data, it also includes many of the actual instruments students could use as a starting point if they wish to do some polling for the 2010 or 2012 elections.

It is not our place to comment in any depth about what will replace *STATS*, though what we have heard suggests we—whether students or practitioners—will be pleased.

Fritz Scheuren


Have you ever purchased a
Which of the following categories best describes your
How many times in the
Which of the following categories best describes
What makes this a quality
NOVEMBER
In the last three months
Alabama
45
Never
yes
Graduate School
sometimes
18-35
monthly

A QUICK GUIDE to Online Data Quality:

Ensuring High-Quality Data from Panel Research

by Vincent Lampone

Online surveys are increasingly popular, and the behavior of market research professionals suggests they are here to stay. Indeed, nearly half the quantitative research in the United States is being conducted through online panels, according to a Forrester Research white paper titled "Is the Long Online Panel Quality Nightmare Over?" Given the limitations of other modes—especially telephone, with its falling response rates—one can easily see why. In short, online panels have allowed survey managers to do more with less.

As the reliance of these panels continues to increase, though, so does the agonizing and hand-wringing among the market research community. There is an increasing fear, whether justified or not, that the panels are inherently unrepresentative, with opaque procedures for quality management that can put companies at risk for bad decisions. This fear is magnified by many market research departments shrinking considerably and being forced to outsource the heavy lifting of quality management to these same vendors.

As such, there is a nascent movement afoot to take back control of data quality on the Internet. Survey managers are attempting to compensate for lost time by establishing a common set of standards for maximizing the quality of their online panel data. This is as true for my organization, National Public Radio, as any other commercial enterprise. We rely heavily on online data collection to support our work. Instead of being passive consumers of research, we need to actively verify vendors' output to have a high degree of confidence in our analysis.

To successfully lead a survey research project, one must be actively engaged in all aspects of the process, from conception to reporting. Here, I offer an overview of the unique issues involved in exercising such leadership for online panel surveys. I synthesize a wealth of research on the topic and aim to offer practical suggestions for how managers can improve data quality.

Panel Research Is Not Always the Right Approach

It is important to begin with a clear statement: Online panel research is not always the right methodology for market researchers, despite its many advantages.

To be sure, online panel research has become popular for many good reasons. More times than not, it is less expensive than telephone surveys and

can be executed in a much shorter time. Moreover, online surveys can include video, audio, and interactive images along with text. Furthermore, it is undeniable that countless organizations—including mine—have used online panel research successfully to guide business decisions and have seen minimal or manageable differences between the results of their online and telephone surveys. Randy Brooks offers an excellent summary of this point in a recent *Quirks* article titled "Internet Data Quality." His main point can be summarized as follows: "Let's not self-flagellate too much, my colleagues. We've been using online data for years, and with commonsense improvements, we will continue to have success in doing so."

Even so, it is imperative that one fully understands the drawbacks of online panels before using them. First and foremost, they are not truly random—nearly all are composed of people who have opted to participate in regular online research. These respondents make up a distinct minority of the American population, and they may or may not represent the opinions or have the same characteristics of the population. For example, those without Internet access are, by definition, not included, and the very poor and very wealthy are under-represented.

In short, when precision and projectability are paramount and cost is less of a concern, other modes of research will continue to be more appropriate than online panel surveys. The latter should only be used with care and after much consideration of alternatives.

Key Obstacles to Data Quality in Online Panel Research

All modes of research have their unique challenges in terms of data quality. For online panel surveys, the largest obstacle is fraudulent or inattentive survey-takers—respondents who, either intentionally or by accident, compromise the integrity of the data set by giving inaccurate answers or not paying sufficient attention to their responses.

This broad category includes many kinds of survey-takers, some posing more of a problem than others. As online panel research is much newer than its counterparts in the survey research field, professionals are still trying to understand the relative impact of each breed of 'problem respondent.' Nevertheless, a brief rundown of each follows, along with a description of why they are harmful.

Fraudulent respondents

These survey-takers intentionally misrepresent themselves or provide inaccurate information. Their goal is usually to maximize the incentives they earn. Research suggests they are, by far, the biggest obstacles to online data quality. Even if their overall participation in individual surveys may not be large, managers should go through great pains to eliminate them from their samples. Unfortunately, it can be somewhat difficult to do so, as they are actively trying to avoid detection.

'Professional' respondents

These survey-takers are not fraudulent, per se, but they do take a lot of surveys. They typically belong to multiple online panels and take at least 10 to 15 online surveys a month. It is still unclear how much these 'professional' respondents influence data quality—some studies have found statistically significant differences between their answers and those of other respondents, while others fail to note many differences. Still, there seems to be an emerging consensus that no harm is done by eliminating these respondents in the sample, and that data quality is likely to improve as a result.

Inattentive respondents

These survey-takers simply do not pay enough attention to their responses. Both fraudulent and 'professional' respondents can fall into this category, but there are many other potential reasons for respondents to become inattentive. For instance, consider long surveys on boring topics, with multiple grid questions and exhausting lists of attributes. These can cause even the best-intentioned respondent to lose interest. Other times, respondents may simply lack the time to give a survey their best effort. This, in turn, can cause phenomena such as "speeding" (racing through a survey as quickly as possible) or "straightlining" (checking the middle option for all choices in a grid without actually reading the questions).

In the end, all survey managers want a sample that is as representative of their target population as possible. They also want respondents who thoughtfully and completely answer each question. Fraudulent and inattentive survey-takers compromise the manager's ability to achieve these goals. So, efforts to improve online quality must focus on severely reducing their prevalence in the sample.

Improving Data Quality at Each Stage of the Survey Process

Throughout the entire survey process, managers can exercise leadership by making the improvement of online data quality their highest priority. This can be done by ensuring all panel partners adhere to the highest standards, designing online surveys that fully engage respondents, and keeping all 'questionable' data out of the final data set. Below is a set of best practices, culled from the latest research from across the industry:

> "In all survey research, not just those conducted on-line, high-quality data is contingent on high-quality surveys. All the same rules of good questionnaire design apply."

The first part: selecting the right panel vendor

There are many online panels out there. Not all are created equal. Size and reputation are important factors to consider, but managers have been hampered by a lack of transparency in many vendors' practices and by a lack of objective criteria through which to evaluate the players in the market.

This is slowly changing, however. In August of 2005, the European Society for Opinion and Marketing Research (ESOMAR) released its "25 Questions to Help Research Buyers." True to its billing, this was—and is—a comprehensive list of questions research customers can ask of panel vendors to shine a spotlight on the quality of their sample. More recent advances include RFL Communications' newly published guide, "Platforms for Data Quality Progress," and the Advertising Research Foundation's ongoing Data Quality Initiative.

Looking at these reports in conjunction, it becomes clear that managers must demand the highest standards of all panel partners in the following three areas:

1 Acquiring panelists intelligently and verifying their identities

It matters how research companies gather folks to participate in their panels. Some participants are recruited by the panels. Others come from "aggregators," outside companies that do their own recruiting and sell samples to many panel owners. And many vendors allow any interested individual to immediately join their panel via the Internet.

In general, the following principles are true:

Invite-only policies are preferable to those that allow anyone to join

Recruitment done by the panel is preferable to that done by aggregators

Panels that rely heavily on aggregators and web recruitment are more likely to attract 'professional' respondents who are solely motivated by incentives, which invites fraudulent and/or inattentive survey-taking.

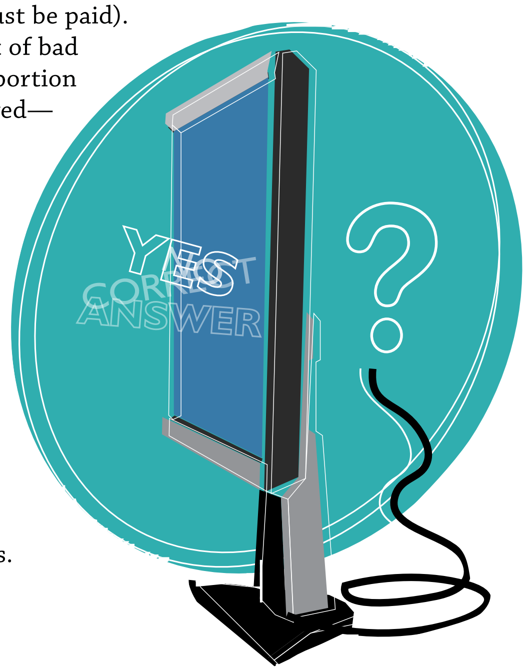
The other piece of the puzzle is in verifying the identities of panelists. This is to prove they are, indeed, actual people and not insidiously represented multiple times in the same panel (under different names or aliases).

Until recently, this was nearly impossible to accomplish. However, with the advent of two new technology solutions, panels and their customers now have options. The most promising of these is called TrueSample, which uses a third party to test sample identities against publicly available sources. People whose information cannot be verified through these sources are deleted from the sample. Furthermore, "virtual machine fingerprinting" is used to ensure the same computer or machine does not exist multiple times in the same database.

This verification process is not currently standard practice for most panels. It also costs more to implement, as a sample must be run through a third party (which must be paid). However, given the obvious cost of bad data—and potentially large proportion of panel sample that can be flawed—the investment is worthwhile.

2 Treating panelists like gold

Panelists are an extremely precious resource; unsatisfied participants can easily jump ship, never to return. Yet, too many panels mistreat this precious resource, pursuing short-term profit at the expense of data quality and sustainability. Survey managers should clearly avoid such vendors.



You can tell if a company is seriously invested in its panelists by asking about its panel management policies, such as the following:

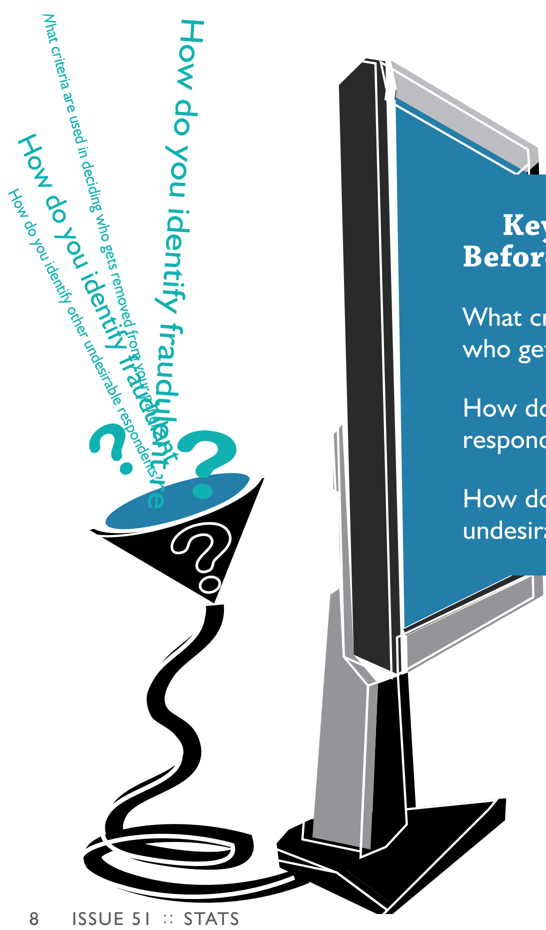
Panelist communication—What is the maximum number of surveys for which a respondent can receive invitations each month? How many times are they emailed for each survey? Responsible companies should not overburden their panelists.

Responsible time limits—Respondent cooperation can fall precipitously after 20 minutes. How does the panel deal with customers who ask for surveys that run over this period?

Questionnaire quality—Does the team have a trained survey expert who reviews all questionnaires for quality (including spelling, grammar, logic, and respondent burden), or does it field all surveys as requested?

Respondent support—Does the panel have a help desk or provide other support to respondents as they take a survey?

Satisfaction—Does the panel collect information about how satisfied respondents are with each survey they take? If so, is this information shared with the customer?



Key Questions to Ask Before Selecting a Vendor

What criteria are used in deciding who gets removed from your panel?

How do you identify fraudulent respondents?

How do you identify other undesirable respondents?

Incentives—While incentives that are too attractive risk unwanted attention from fraudulent respondents, they should be sufficient for attracting and retaining good panelists. What incentives are panelists offered?

3 Strenuously weeding out undesirable respondents, in all their forms

Fraudulent, inattentive, and 'professional' respondents all degrade survey data. Panels need to actively monitor the quality of their sample and remove panelists who consistently commit fraud or become disengaged.

The trick, however, is in deciding how stringent the criteria should be in determining which respondents should be booted. Sometimes, respondents have reason to be disengaged from surveys that, for whatever reason, don't hold their interest. Different companies are likely to have varying interpretations of who constitutes a 'problem' respondent. It is critical for survey managers to understand what definitions are used by each company and determine their level of comfort before selecting a vendor. Key questions to ask include the following:

What criteria are used in deciding who gets removed from your panel?

How do you identify fraudulent respondents?

How do you identify other undesirable respondents?

The second part: designing the survey research strategically

Selecting the right panel vendor is important, but it is clear survey managers should not completely delegate their responsibility for data quality to the panels.

Quite a bit of research has been done on techniques online survey designers can implement to improve data quality. Fortunately, there are many at the manager's disposal, such as the following:

Including respondent 'traps' in the survey instrument—This practice is increasingly common among market researchers. It involves the judicious use of traps to record the engagement of survey respondents. Those who fail these traps are likely to be fraudulent or inattentive, and the survey manager may consider removing them from the sample. To compensate, the manager must over-sample at the outset—10% to 15%, according to Theo Downes-LeGuin's white paper titled "Satisficing Behavior in Online Panelists"—so he or she still ends up with a desired/required number of completes.

In a November 2007 article in *Quirks*, Kurt Knapton and Rick Garlick provide an excellent summary of four of the most commonly used traps, including the following:

Red herrings—providing a fake brand, service, or characteristic among a list and removing respondents who select it (e.g., those who claim to have listened to the invented NPR show "From Milan to Minsk" in the past week).

Consistency of answer traps—asking the same fact-based question at the beginning and end of a survey and removing respondents who answer differently (e.g., those who claim their age is 39 and say they were born in 1953).

Mutually exclusive traps—including questions with answer choices that are oppositely worded and removing respondents who aren't consistent (e.g., those who say they liked a program on NPR and also disliked the same program).

Simple instruction traps—asking respondents to enter a specific response and removing those who fail to do so (e.g., "please check slightly disagree," "please enter the number 24 here").

Also, to measure the prevalence of 'professional' responders in the sample, the survey manager can include a question asking participants how many panels they belong to or how many online surveys they have taken in the past month. Those who exceed a certain threshold may then be removed.

Finally, on the back end, survey managers can account for the following two phenomena when conducting data analysis:

Straightlining—detecting response patterns that are too predictable to be credible

Speeding—detecting respondents who complete the survey in an unreasonably short period of time

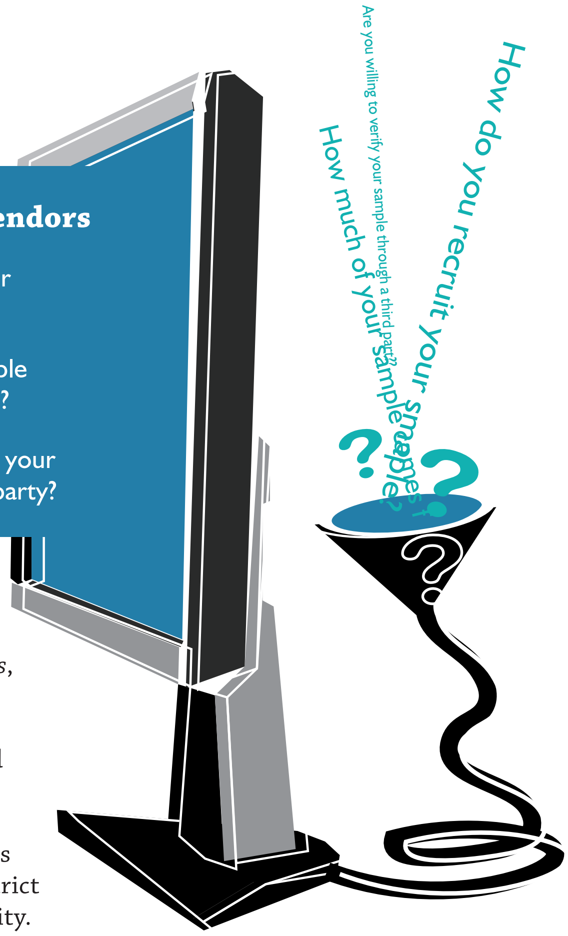
All these techniques can prove useful to the survey manager in the right context in identifying disengaged respondents. Nevertheless, one should avoid a 'kitchen sink' approach that makes a survey overly burdensome. Also, these traps are not fool-proof—many are easily avoided by professional responders, who are quick to adapt to researchers' efforts, and some (like mutually exclusive traps) may end up nabbing too many innocent people. As usual, discretion and good sense are invaluable guides.

Questions to Ask Vendors

How do you recruit your sample?

How much of your sample comes from aggregators?

Are you willing to verify your sample through a third party?



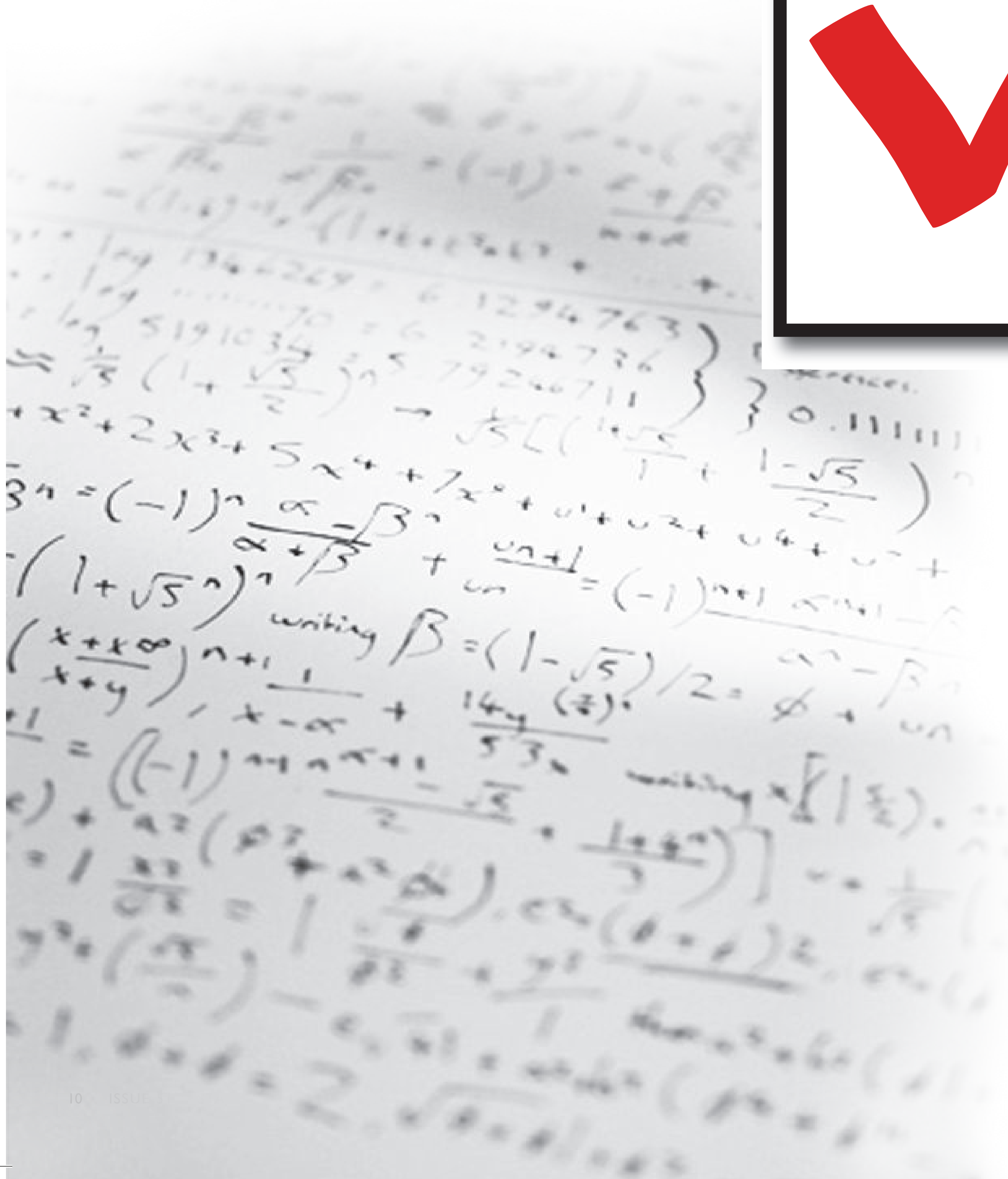
In a recent issue of *Quirks*, Jon Puleston and Deborah Sleep called boredom "the survey killer"—and for good reason. Fortunately, the industry, in its pursuit of improved data quality, seems to be moving away from a strict "blame the panelist" mentality. In all survey research, not just that conducted online, high-quality data is contingent on high-quality surveys. All the same rules of good questionnaire design apply. Especially when the survey topic is not particularly fascinating, managers must keep the respondent engaged, have realistic expectations, and carefully review and test instruments to reduce all possible sources of nonsampling error.

Ultimate Responsibility

In the fast-paced environment of most modern organizations, when resources are limited and work demands unyielding, strict attention to data quality can seem like a luxury—a "nice-to-have" quality, but impractical and easy to outsource. An incorrect business decision based on faulty data, however, does lasting damage to both the bottom line and researchers' reputations. For those relying on online panels, it ultimately lies in the hands of the survey manager—not the third-party vendor—to ensure the highest quality data. With the tools outlined here, it is hoped that survey managers are armed with information to assume more effective leadership of data collected by outside panel partners. ◕

Editor's Note: *Data collection in the 21st century is quickly taking on new methods that are going to take getting used to. But they are fun, too, and when approached with common sense, can lead to sound decisions.*

ACCEPT



REJECT…, OR GET MORE DATA?

An Intuitive View of Sequential Tests

by Philip Wong, Roanna Gee, and Bruce E. Trumbo

A pharmaceutical company uses bacteria to 'grow' a protein that is a key ingredient in one of its drugs. Through careful quality management, bioengineers at the company's American plant have stabilized the production of the protein so the yield from batches is normally distributed with mean $\mu = 100$ and standard deviation $\sigma = 20$. However, the company's new European plant gets better yields: normally distributed with mean $\mu = 110$ and standard deviation $\sigma = 20$. Engineers aren't sure whether the higher yields in Europe result from a slightly different formula used there to make batches or from the newer equipment installed there. They want to test the European formula in the American plant to see what happens.

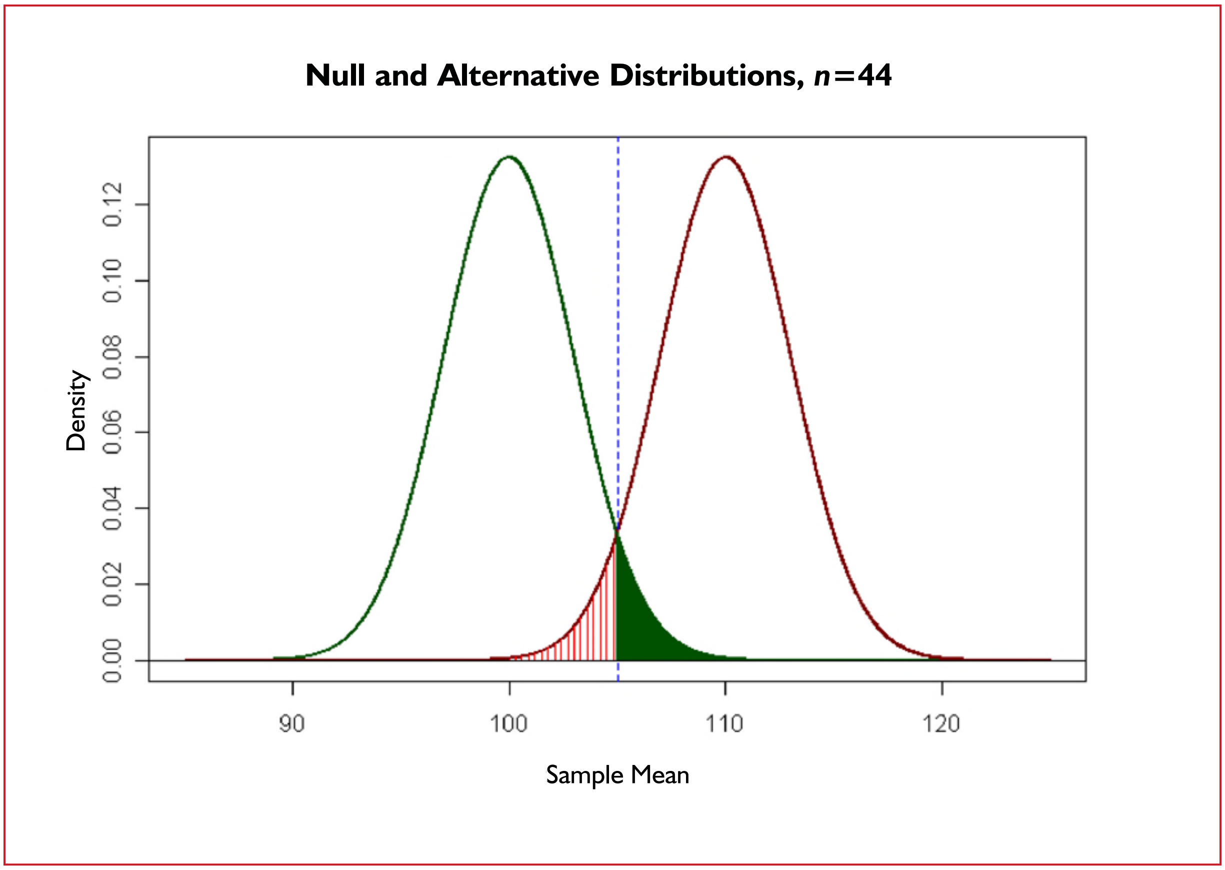


FIGURE 1. Distributions of the sample mean under H_0: $\mu = 100$ (left) and H_1: $\mu = 110$ for $n = 44$ batches. Error probabilities, indicated by shading, are $\alpha = 5\%$ (dark solid area to the right of 105) and $\beta = 5\%$ (light vertical stripes).

Their null hypothesis, H_0: $\mu = 100$, is that the new formula does not change the yield at the American facility. Their alternative, H_1: $\mu = 110$, is that the formula increases the yield to match what the European plant is getting. This testing situation is called "simple against simple," because the null hypothesis and alternative against which it is tested each specify only a single value of the parameter μ.

How Many Batches?

As you might expect, the sample mean yield $\overline{X}$ of some number n of batches made at the American plant with the new formula will be used to decide whether to accept or reject the null hypothesis H_0. Suppose the engineers want only a 5% chance of making a wrong decision. How many batches do they need to run to have a good chance of making the right decision? Specifically, suppose they want

$$\alpha = P(\text{Reject } H_0 \mid \mu = 100) = 5\%$$

and

$$\beta = P(\text{Accept } H_0 \mid \mu = 110) = 5\%.$$

Because α and β are equal, it is not difficult to show that we should 'split the difference' between $\mu_0 = 100$ and $\mu_1 = 110$, accepting H_0 if $\overline{X} \leq 105$ and rejecting H_0 if $\overline{X} > 105$. The value 105, called the critical value, separates values of $\overline{X}$ that lead to accepting H_0 from those that lead to rejecting.

If H_0 is true, then $\overline{X}$ is normally distributed with mean 100 and standard deviation $20/\sqrt{n}$. Standardizing $\overline{X}$, we get

$$\begin{aligned}\alpha &= P\{\overline{X} > 105\} \\ &= P\{Z > \sqrt{n}\,(105 - 100)/20 = 0.25\sqrt{n}\}.\end{aligned}$$

Because $\alpha = 5\%$, we have $0.25\sqrt{n} = 1.645$ from tables of the standard normal distribution. Thus, we need to use $n = 44$ batches. We get the same answer with the following R code:

```
n = 1:100
se = 20/sqrt(n)
alpha = 1 - pnorm(105, 100, se)
min(n[alpha <= .05])
```

Similarly, if H_1 is true, then $\overline{X} \sim \text{NORM}(110, 20/\sqrt{n})$ and again we need $n = 44$ batches to get $\beta = P\{\overline{X} \leq 105\} = 5\%$. Shaded regions in Figure 1 correspond to α and β.

Visualizing the Variability of Sample Means

Ordinarily, we would look at the value of $\overline{X}$ only after all $n = 44$ observations are obtained and then make our decision based on whether $\overline{X}$ is above the critical value of 105 (reject H_0) or below 105 (accept H_0). However, we're interested in how $\overline{X}$ behaves sequentially as we take each of the 44 observations in turn.

The Law of Large Numbers says that, as the sample size n increases, the sample mean $\overline{X}$ converges to the mean of the corresponding population—either 100 or 110 here. But, this tells us what happens as n approaches infinity, and 44 is a long way from infinity. It is useful to look at the trace of $\overline{X}$ as we progress toward $n = 44$. For each value of n from 1 through 44, we plot the average of the first n observations against n.

For example, suppose the alternative H_1: $\mu = 110$ is true. As we take more and more observations, how rapidly and consistently can we expect $\overline{X}$ to stabilize on the high side of the critical value 105, thus leading us to a correct decision? If we simulate this experiment many times, we see the answer can vary considerably from one run to the next:

— Often, as in figures 3(a) and 3(b), the trace establishes itself above 105 early on and stays there until the 44th batch. So, we feel secure in the decision to reject H_0, perhaps even before seeing all 44 batches. As an extreme example, about 26% of such random traces never go below 106 during the 44 steps on the way to rejecting the null hypothesis. (Figure 2 shows the R code we used to simulate this percentage and discover that more than half of the traces stay above 106 for at least 40 of the 44 steps.)

— Sometimes, the trace wanders near the critical value 105 and happens to lie above it when $n = 44$, as in Figure 3(c). Again, we make the correct decision—but maybe feeling uneasy and wishing for stronger evidence. Such close calls are not rare; about 17% of the random traces lead to rejection after staying between 103 and 107 for at least 15 of the 44 steps. (A slight modification of the R code in Figure 2 can be used to find this percentage.)

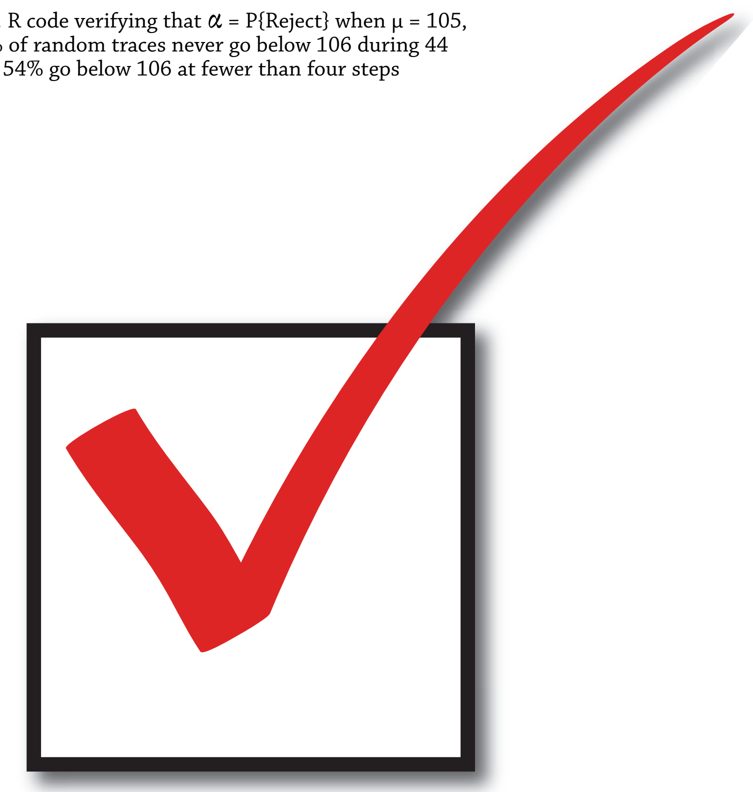
— Of course, $\beta = 5\%$ of the time, the trace will end below 105 by chance, as in Figure 3(d), thus leading us incorrectly to accept H_0. (The value `mean(acc)` in Figure 2 verifies this.)

Simulations can be done in the same way for the case in which H_0 is true ($\mu = 100$). Thus, we can verify that $1 - \alpha = 95\%$ of the traces end below 105 at step 44, leading us to accept H_0.

```
m = 100000     # iterations
n = 44;   mu = 110;   sg = 20
acc = hi = numeric(m)
   # all 0's at start
for (i in 1:m) {
 x = rnorm(n, mu, sg)
   # 44 observations
 av = cumsum(x)/ (1:n)
   # 44 running means
 if (av[44] <= 105) acc[i] = 1
 hi[i] = sum(av > 106)  }
   # hi is number of steps
   #  for which av > 106
mean(acc)
mean(hi==44)
mean(hi >= 40)

[1] 0.04936
[1] 0.25874
[1] 0.54007
```

FIGURE 2. R code verifying that $\alpha = P\{\text{Reject}\}$ when $\mu = 105$, about 26% of random traces never go below 106 during 44 steps, and 54% go below 106 at fewer than four steps



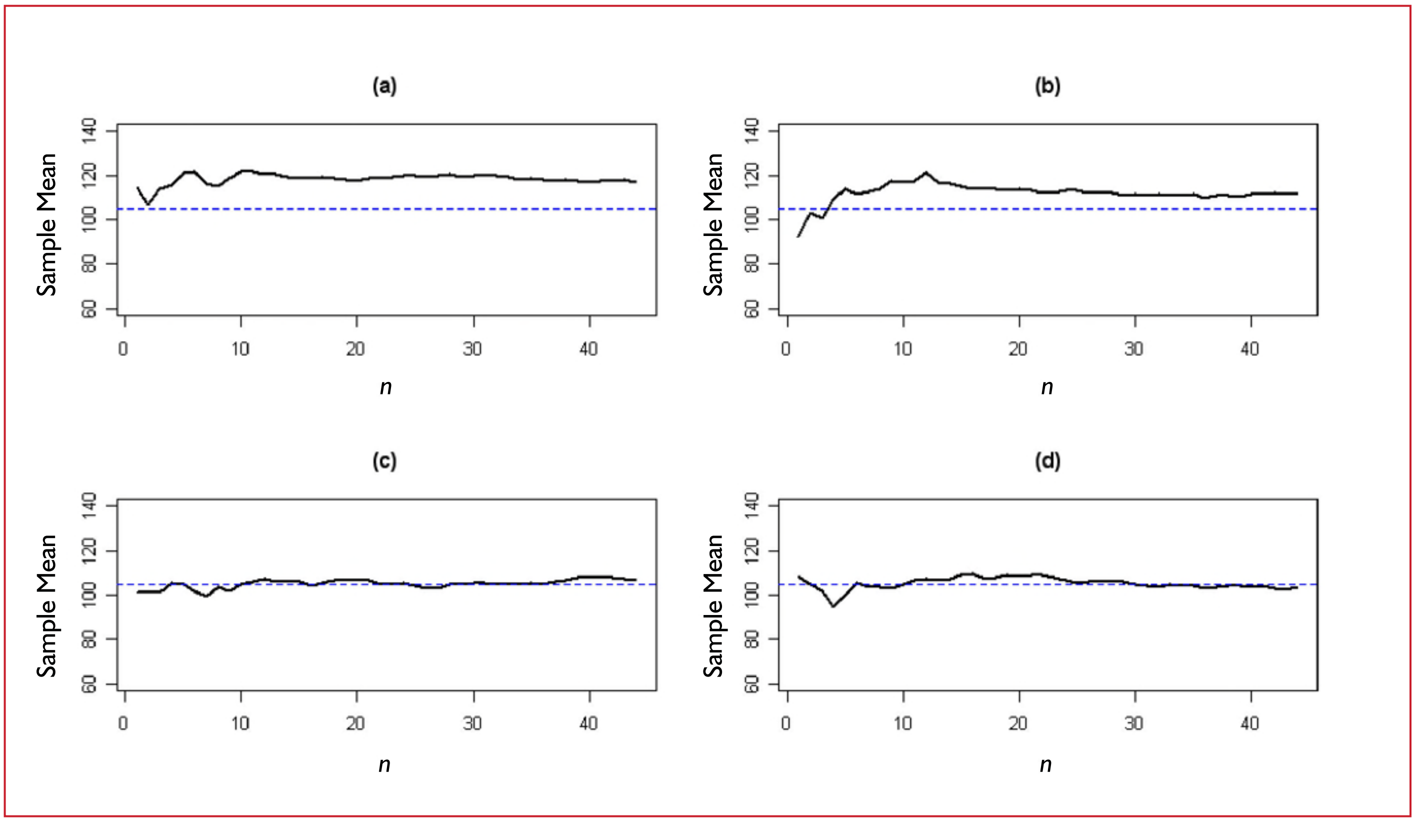


FIGURE 3. Traces of four experiments simulated under H_1: μ = 110, each with n = 44 batches. The broken horizontal line is the critical value 105 with $\alpha = \beta$ = 5% at n = 44. Traces (a) and (b) seem to establish early that the null hypothesis H_0: μ = 100 should be rejected. For trace (c), this decision is not evident as soon. As in (d), 5% of traces in 44 batches lead to the wrong decision, accepting H_0: μ = 100.

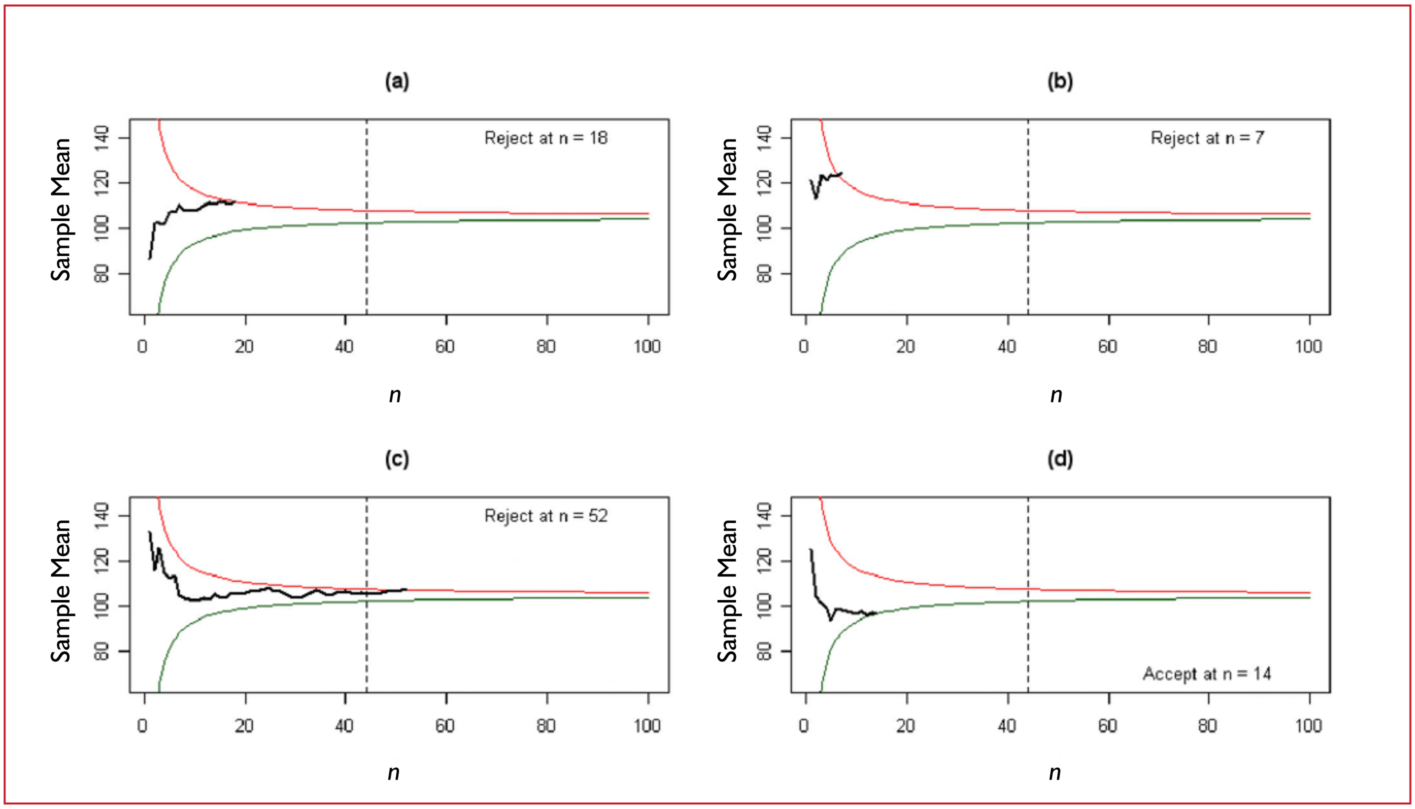


FIGURE 4. Traces of four simulated sequential experiments under H_1: μ = 110. Target error probabilities $A = B$ = 5%. Traces (a) and (b) reach correct decisions with many fewer than 44 batches, which would be required in the corresponding nonsequential experiment. Relatively rarely, as in (c), the experiment requires more than 44 batches before its trace crosses the upper boundary to reject H_0. As in (d), about β = 4% of traces cross the lower boundary, leading to an incorrect decision.

Sequential Sampling and Testing

In the early 1940s, during World War II, the distinguished statistician Abraham Wald had an idea about how it might be possible to reduce the cost of experimentation on newly developed naval ordinance. His idea was to take advantage of situations in which early observations already show clearly whether the null hypothesis will be accepted or rejected. In terms of our problem, the idea would be to look at $\overline{X}$ after each observation, deciding according to some rule whether we have enough evidence to make a decision. If so, end the experiment. If not, keep collecting data until the correct decision becomes clear.

One difficulty with this sequential approach is that the probability theory is too messy to handle analytically. This is because one must deal with sequences of outcomes that might end at any time, not just the two distributions shown in Figure 1. However, Wald was able to find an approximate, and relatively simple, solution to the probability distribution problem. Even so, his argument is a little too complicated to present here, and we just describe how his results are used. (Many books on mathematical statistics show the derivation.)

Wald's procedure is to select 'target' values A and B for α and β and, from these, obtain upper and lower boundaries for $\overline{X}$ at each step in the sequence. For our problem, the boundaries at step n are

$$(\mu_0 + \mu_1)/2 \pm 2.944\sigma^2 / (\mu_0 - \mu_1)n,$$

where the constant 2.944 is derived from $A = B =$ 5% as ln(0.95/.05). Notice that $(\mu_0 + \mu_1)/2 = 105$, so these upper and lower hyperbolic boundaries are located symmetrically about the critical value 105 we used in the nonsequential test:

If the trace of $\overline{X}$ crosses above the upper boundary at a step, we stop the experiment there and reject H_0.

If the trace of $\overline{X}$ crosses below the lower boundary at a step, we stop the experiment there and accept H_0.

As long as the trace of $\overline{X}$ stays between the boundaries, we keep looking at more batches.

Wald's approximate procedure guarantees that the sum of actual error probabilities $\alpha + \beta$ will not exceed the sum $A + B$ of the target probabilities. But, his procedure does not specify the actual values α and β.

Figure 4 shows Wald's boundaries and how they work for four simulated runs of our experiment. The observations are simulated according to the distribution NORM(110, 20) so that H_1 is true and our traces will usually cross the upper boundary, rejecting H_0.

A MILITARY SECRET REVEALED

Although Abraham Wald developed sequential testing in the early 1940s, publication of his first public paper on the topic was delayed until 1945—the end of World War II—because the National Defense Research Group "considered these developments sufficiently useful for the war effort to make it desirable to keep the results out of the reach of the enemy, at least for a certain period of time."

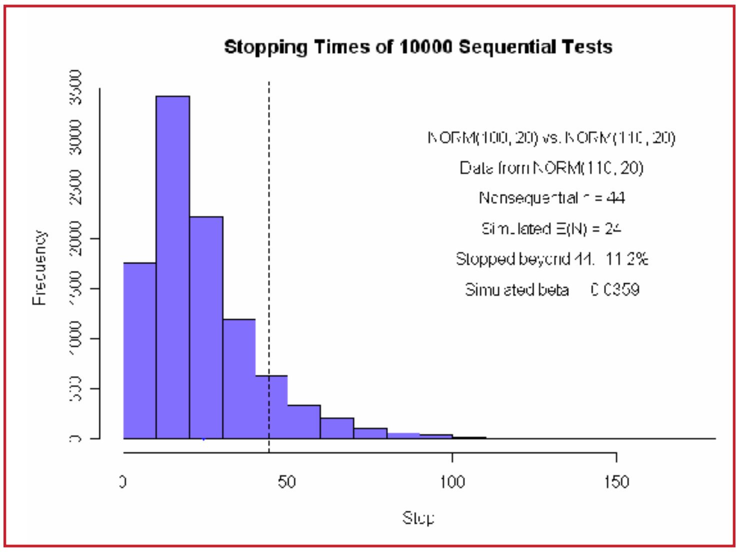


FIGURE 5. The simulated distribution of stopping times N. The mean sample size 24 is considerably smaller than the 44 batches needed for a nonsequential experiment. Achieved $\beta \approx 4\%$.

Often, the process ends after we look at only a few batches. Sometimes, however, it continues to large sample sizes. And, of course, it occasionally results in a wrong decision. Fortunately, traces with quick and correct decisions predominate. We had to do a lot of simulation runs to get a panel of four figures in Figure 4 that illustrate a variety of outcomes. (Traces such as 4(c), and especially 4(d), are relatively rare.) Because we chose $A = B$, similar runs using $\mu = 100$ behave symmetrically, with the process usually crossing the lower boundary and ending in acceptance of H_0.

When the trace crosses a boundary, ending the experiment, it typically goes at least a small distance beyond the boundary. A noticeable 'overshoot' is visible in Figure 4(b). We mention this because Wald's approximation makes the simplifying assumption that all traces end exactly at the boundary. Overshoots are a major reason why the target values A and B are not achieved exactly using his formulas for boundaries.

Simulating the True Properties of a Sequential Test

Important questions arise in a sequential procedure that can be answered best with simulation. For a sequential test, the sample size N is a random variable. One never knows in advance exactly when the procedure will end.

How big might N be? On average, the sample size N is smaller than the sample size n that would be required in an experiment with a fixed sample size. In our example, $E(N) < n = 44$. Specifically, with normal data and the target error values $A = B = 5\%$, Wald's approximation is $E(N) \approx 21.2$. In any one experiment, however, there is a small probability that N might be much larger than the fixed sample size for a nonsequential test.

Figure 5 shows a histogram of the values of stopping times N in 10,000 simulated sequential experiments. We see that the actual value of $E(N)$ is about 24. This is still a substantial savings, compared with $n = 44$ for the nonsequential test. Even though about 11% of the sequential tests require more than 44 batches, the simulation also shows that few require more than 100 batches. We also see from this simulation that the actual error probability β is roughly 4%, which is below of the target value $B = 5\%$.

So, the actual mean sample size is a little larger (worse) than Wald's approximation says, and the error probabilities are a little smaller (better) than the target values. For many purposes, these actual values might be as satisfactory as the anticipated ones. But, we could come closer to $E(N) = 21.2$ and closer to $\alpha = \beta = 5\%$. We could do this by adjusting the boundaries for acceptance and rejection so they are closer—so that continuations are slightly less common. For example, we tried target values $A = B = 6\%$, which changed the constant in the boundary equations from 2.944 to 2.759 and resulted in actual error probabilities a little below 5% and $E(N)$ about 22.

As a result of the symmetry from choosing $A = B$, a simulation under H_0: $\mu = 100$ shows the distribution of N to be essentially the same as in Figure 5 and α at about 4%. When $A \neq B$, the distribution of N depends on whether H_0 or H_1 is true. In particular, $E(N \mid H_0)$ typically differs from $E(N \mid H_1)$. Because μ is unknown, we wouldn't know which of these two expected sample sizes is correct, but we could assume the worst and use the larger value for planning purposes.

In practice, we might end the process artificially at some predetermined point, say 100 batches, if the trace hasn't already crossed a boundary by then. In that case, we would 'truncate' the experiment at step 100, rejecting or accepting depending on whether $\bar{X}$ is larger or smaller than 105, respectively. Runs beyond $n = 100$ are so rare that forced stopping at 100 has little effect on the error probabilities or expected sample size. The R code for truncating a test at 100 is shown in Figure 6. This program is simple, but a bit wasteful; for every experiment, we simulate 100 batches and ignore the batches not needed. (Wasting simulated batches is a lot cheaper than wasting real ones.)

```
m = 100000;   t = 100;   n = 1:t
sg = 20;    mu = 110
up = 105 + 2.9444*sg^2/(10*n)
lw = 105 - 2.9444*sg^2/(10*n)
rej = acc = stop = numeric(m)
for (i in 1:m)  {
  x = rnorm(t, mu, sg)
  av = cumsum(x)/n
  N = min(n[(av >=up)|(av <=lw)], t)
  stop[i] = N
  rej[i] = (av[N] >= up[N])
  acc[i] = (av[N] <= lw[N])   }
mean(stop); mean(acc); mean(stop==t)

[1] 24.08002
[1] 0.03779
[1] 0.00399
```

FIGURE 6. If a sequential test of H_0: $\mu = 100$ against H_1: $\mu = 110$ has $\sigma = 20$, target error probabilities 5%, and truncation at 100 steps, then the mean stopping time is about 24 and the achieved error probability $\beta = P\{\text{Accept} \mid H_1\} \approx 4\%$. Actual truncation is rare.

Uses of Sequential Procedures

Sequential testing has recently become much more widely used than in the years immediately following its invention. Computer simulation has made feasible what cannot be done by analytic methods. We have used simulation to get good approximations of the exact properties of a sequential test—the achieved error probabilities α and β and facts about the distribution of N, for example.

Sequential tests are most commonly used in fields where ethical or financial concerns make it especially important to end testing as soon as a decision can reliably be made. Clinical trials of new drugs provide important examples of situations in which sequential tests may be helpful. If drugs to cure or alleviate a serious illness are under test, researchers managing a clinical trial want to end it as soon as possible so all subjects can be treated with the drug that proves most effective. ◕

Author's Note: *Over the last dozen or so years, it has been my privilege to write regularly for STATS. With this final issue, I would like to thank the editors, reviewers, production staff, and fellow authors—as well as my student 'referees' and co-authors—for their creativity and hard work. The result has been many excellent issues of great benefit to statistics students and the profession. - B.E.T.*

CHALLENGES

Annotated computer code for all figures is available at *www.amstat.org/publications/stats*.

1. The simulation for Figure 5 shows that β is about 4%, while a similar simulation shows that α is also about 4%. It seems fair to compare $E(N) = 24$ for our sequential test with a nonsequential test based on a fixed sample size n chosen to give $\alpha = \beta = 4\%$. What is the required value of n?

2. Even if $\alpha \neq \beta$, the sample size of a nonsequential test can be found as $n = [(z_\alpha + z_\beta)\sigma/\Delta]^2$, where z_α and z_β cut off probabilities α and β, respectively, from the upper tail of a standard normal distribution and Δ is the difference between μ_0 and μ_1. Verify that this formula gives $n = 44$ when $\alpha = \beta = 5\%$ for the example used in this article. What sample size is necessary when $\sigma = 20$ and $\Delta = 10$ and the desired error probabilities are $\alpha = 1\%$ and $\beta = 5\%$? What is the critical value in this case?

3. Modify the code in Figure 2 for $\alpha = \beta = 4\%$, do the simulation, and report your findings.

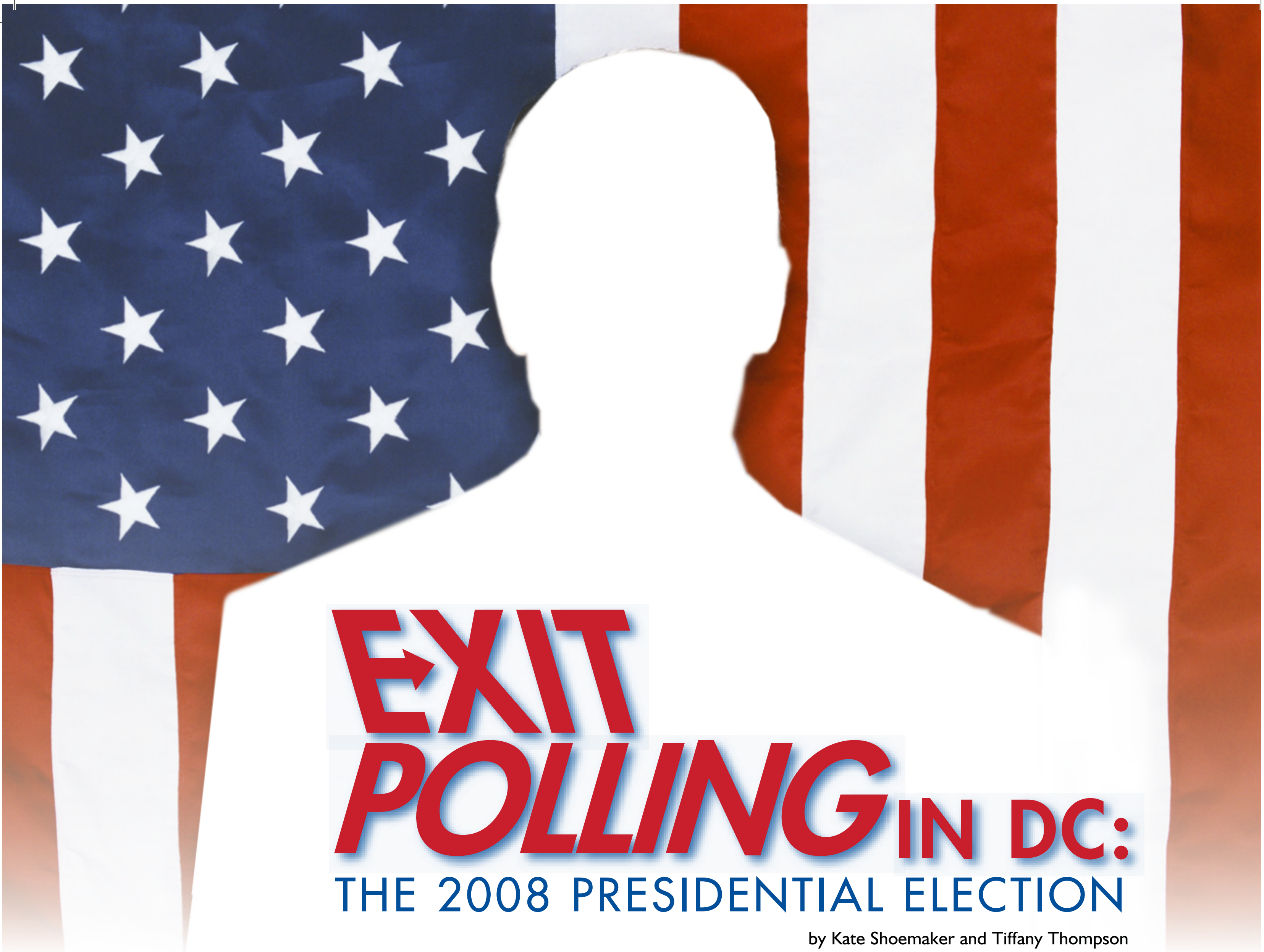
4. Modify the code in Figure 6 and give results for a sequential test truncated at step 80.

5. Use the R code at *www.amstat.org/publications/stats* to make your own version of Figure 5. Modify this code to find the actual values of α and $E(N \mid H_0)$ for when H_0 is true.

6. (Advanced) Sometimes it is more economical to look at $\bar{X}$ to make a decision only after each group of g observations. Intuitively, what effect would this have on $E(N)$ and the achieved error probabilities α and β? Modify our R code to give quantitative answers for $g = 5$, $\mu_0 = 100$, $\mu_1 = 110$, $\sigma = 20$, and $A = B = 5\%$.

To check your answers, visit the *STATS* web site at *www.amstat.org/publications/stats*.

About the Student Authors: At the time this article was written, both Roanna Gee and Philip Wong were MS students in statistics at California State University, East Bay. Since then, Wong has been working as an insurance analyst for the California State Automobile Association and Gee for Wells Fargo Bank, with a group that plans procedures to minimize operational risk in electronic transactions.



EXIT POLLING IN DC: THE 2008 PRESIDENTIAL ELECTION

by Kate Shoemaker and Tiffany Thompson

On November 4, 2008, Kate Shoemaker and Tiffany Thompson polled a precinct in Washington, DC, from 4:30 p.m. to 7:30 p.m. They found that their sample voted 75% for then Sen. Barack Obama and 19% for Sen. John McCain. While the certified results from the precinct of their polling location showed Obama securing a majority of the vote, Shoemaker and Thompson's results indicated less of a landslide. Shoemaker and Thompson broke down the results by each interviewer, and it appears the participants answered similarly.

Shoemaker and Thompson's goal was to design, conduct, and analyze the results of a poll of Washington, DC, voters. The questions were carefully chosen while keeping the geographical area in mind. Two questions regarding local DC issues were included, and the composition of the polling sample was captured with demographic questions. Shoemaker and Thompson wanted to find out for whom the participants voted, which issues were most important to them, and how they felt about the current political situation. Additionally, they wanted to compare the results of their polling location with the certified results reported by the District of Columbia's Board of Elections and Ethics.

Voting Location

The polling location was in Dupont Circle, which is known for being a culturally diverse and liberal-leaning neighborhood. Two other groups from The George Washington University performed similar polls in Washington, DC, while other students performed polls in northern Virginia and Maryland.

In 2006, *The Washington Post* displayed socio-economic data for the Dupont Circle area using data from Environmental Systems Research Institute (ESRI).

Those in the polled precinct had reliably voted for the Democratic presidential nominee in the past. For the 2004 presidential election, 87% of the vote went to Sen. John Kerry and 9.44% to President George Bush. For the 2000 presidential election, Sen. Al Gore got 75% of the vote, while then Gov. George Bush got 10.5%. Therefore, Shoemaker and Thompson were not surprised to see the Democratic voting trend continue into the 2008 presidential election.

The polling location had both electronic machines and paper ballots, allowing voters to choose their preferred method. The location was open from 7 a.m. until 8 p.m. Voters standing in line at 8 p.m. were allowed to vote, regardless of the time they actually made it inside the polling location.

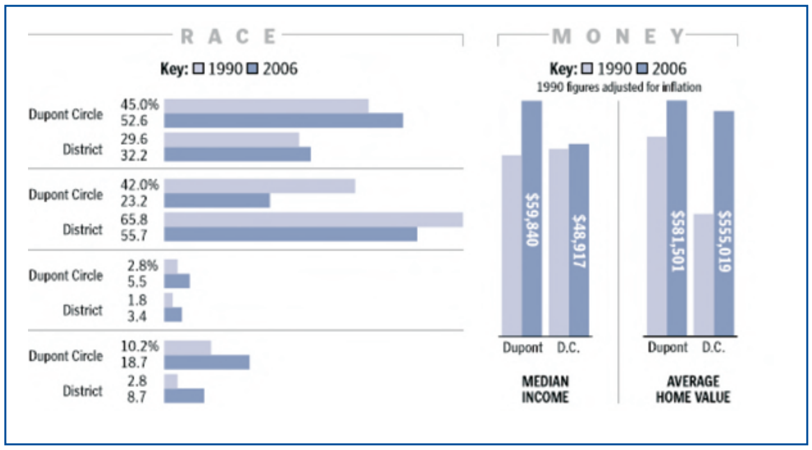


FIGURE 1. Data show the polling area (Dupont Circle in Washington, DC) as becoming increasingly racially diverse. Also, data show higher incomes and home values, compared to the rest of DC.

Collection Methods

Shoemaker and Thompson developed a survey questionnaire in which all questions were carefully selected and phrased. First drafts of the questionnaire were distributed in a pre-test, and alterations were made for the final version. Additionally, two questions dealing with DC citizens' voting rights and the DC public school tuition voucher program were asked.

Shoemaker and Thompson tried to be as prepared as possible for Election Day by creating a general polling observations form and opening and closing voter counts form that were completed during the polling. They predetermined their sample size (a minimum of 30) and calculated that they should attempt to contact every eighth voter. They also were prepared to take down as much information as possible about nonrespondents through an exit poll nonresponse form. Responsibilities were assigned to each team member and potential problems were discussed.

Absentee votes comprised about 9% of the registered voters in the precinct. No attempt was made to poll absentees. The Election Day turnout for the precinct was about 54%.

Election Day

Shoemaker and Thompson arrived at their polling location around 4:30 p.m. on Election Day. There was no one standing in line to enter the building. There were people standing about 50 feet from the polling location, attempting to distribute information supporting their preferred candidate. Shoemaker and Thompson noted it was raining lightly, a little chilly, and getting dark.

After going inside and introducing themselves to the election official, Shoemaker and Thompson noted it was not crowded. The election official told them there was a high turnout early in the morning and again at lunchtime. About 2,500 people had already voted, which the election official estimated to be about double the turnout from the 2004 election. Shoemaker and Thompson were allowed to stand relatively close to the door out of which voters exited.

Arriving at their designated area to set up operations, Shoemaker and Thompson met a graduate student from The George Washington University who was exit polling for the National Election Pool, conducted by Edison Media Research and Mitofsky International. She had been exit polling all day and confirmed that the morning had been more crowded than it was at that time.

Shoemaker and Thompson found exit polling in the rain to be tough, as they juggled umbrellas, clipboards, the survey drop box, and other supplies.

#	Interviewer	Status			Gender		Age			Race/Ethnicity					Notes
1	Kate	Ref	**Miss**	NE	**M**	F	18-34	**35-54**	55+	**W**	B	H	A	DK	
2	Kate	**Ref**	Miss	NE	M	**F**	**18-34**	35-54	55+	**W**	B	H	A	DK	
3	Kate	**Ref**	Miss	NE	M	**F**	**18-34**	35-54	55+	**W**	B	H	A	DK	pregnant
4	Kate	**Ref**	Miss	NE	**M**	F	**18-34**	35-54	55+	W	B	H	A	**DK**	
5	Kate	**Ref**	Miss	NE	**M**	F	18-34	**35-54**	55+	W	B	**H**	A	DK	
6	Tiffany	Ref	**Miss**	NE	M	**F**	18-34	35-54	**55+**	**W**	B	H	A	DK	"still polling?"
7	Tiffany	**Ref**	Miss	NE	M	**F**	18-34	35-54	**55+**	W	B	H	**A**	DK	lady w/ baby
8	Tiffany	**Ref**	Miss	NE	**M**	F	18-34	35-54	**55+**	**W**	B	H	A	DK	rude guy
9	Tiffany	Ref	**Miss**	NE	**M**	F	**18-34**	35-54	55+	**W**	B	H	A	DK	
10	Tiffany	**Ref**	Miss	NE	M	F	18-34	35-54	55+	W	B	H	A	DK	
11	Tiffany	**Ref**	Miss	NE	**M**	F	**18-34**	35-54	55+	**W**	B	H	A	DK	
12	Tiffany	**Ref**	Miss	NE	**M**	F	**18-34**	35-54	55+	**W**	B	H	A	DK	

TABLE 1. Summary of Nonresponse Observational Form

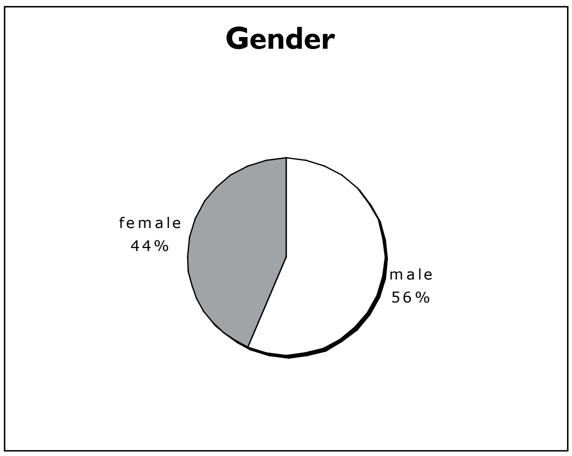


FIGURE 2. Participant demographics

Additionally, the polling location had only one official exit, but some voters mistakenly exited out of the polling entrance. Though they tried to catch these voters, Shoemaker and Thompson noted they may have missed some. Further complicating matters, election workers often exited through the entrance and were initially difficult to differentiate from voters.

For the most part, people were willing to take the survey. Shoemaker and Thompson originally chose to sample every eighth voter, but revised this to every fifth voter at 5:45 p.m., due to not many people coming in to vote. One unforeseen complication was people asking to participate in the exit poll, which could have negatively affected the quality of the data.

Shoemaker and Thompson split the responsibility of approaching voters to fill out the questionnaire and filling out the nonresponse form and holding the ballot box. They both kept track of people exiting, so as to keep track of every nth person. Also, they both noted general observations about the polling process and their surroundings.

It took two and a half hours to complete the exit poll, which was longer than Shoemaker and Thompson anticipated and thought to be due to the low number of people who voted that evening.

Data Analysis

After finishing the survey, Shoemaker and Thompson immediately began discussing the entire survey experience. They went over their observational forms to make sure they had not left out any important information and added notes while the experience was still fresh in their minds. They discovered they had done a thorough job of completing their observational sheets, which was a concern due to the numerous distractions that competed for their attention. They also scanned their completed questionnaires and further discussed their nonresponse observational form (see Table 1).

Shoemaker had five voters who chose to not respond; Thompson had seven. Shoemaker's nonresponding voters were mostly young, white men. However, she did have a woman who was pregnant. Thompson's nonresponding voters also were white men, but were either young or old (not middle-aged). One possible reason as to why Thompson had more refusals than did Shoemaker is that she went second, making it later in the evening when she did her polling.

Shoemaker and Thompson then analyzed their completed questionnaire forms, coding the questions and inputting all responses into an Excel spreadsheet. They were surprised to find that almost all participants completed the entire survey.

Results

Shoemaker and Thompson chose to poll every eighth person for the first half of the survey and every fifth person for the second half. Figure 2 shows the demographics of the participants.

Although both genders were represented, there was a higher percentage of males—18 compared to 14 females. The majority of participants were white, followed by Hispanics and then African Americans and Asians. One person identified with both the multiracial and "other" categories, while no one identified himself or herself as American Indian, Native Hawaiian, or Pacific Islander.

Figure 3 shows that about 84% of the survey respondents were between the ages of 18 and 39. This was expected due to younger people being less likely to have children they need to take care of after work. Figure 4 shows a good representation across income categories, though more than 50% of the participants chose one of the two lowest income categories.

Figure 5 shows that the majority of participants finished at least some postgraduate work. Also, everyone in the sample had completed high school. The majority of the sample self-identified as Democrats, while Independents and Republicans made up 13% (see Figure 6). The two pie charts in Figure 7 show how the sample voted and the certified results for the precinct in which Shoemaker and Thompson took their poll.

Seventy-five percent of the sample voted for Obama, while 90% of the entire precinct voted for him. Nineteen percent of the sample voted for McCain, while 8% of the precinct voted for him.

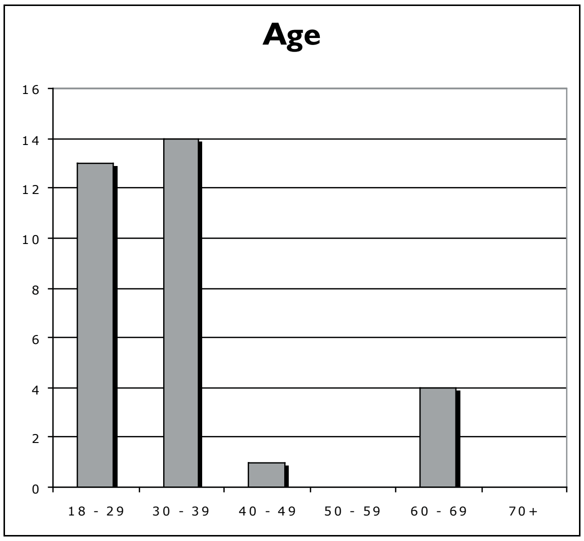


FIGURE 3. Eighty-four percent of respondents were between the ages of 18 and 39.

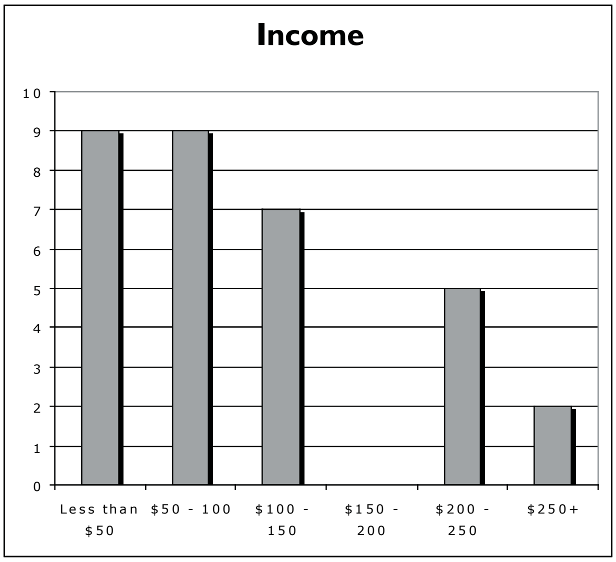


FIGURE 4. Participants' income levels

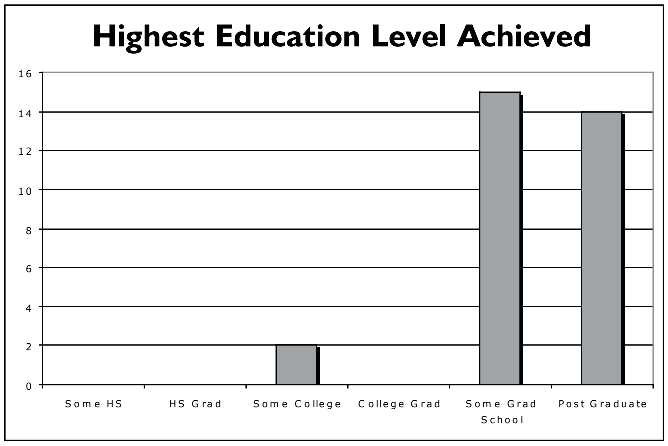


FIGURE 5. The majority of participants finished at least some postgraduate work.

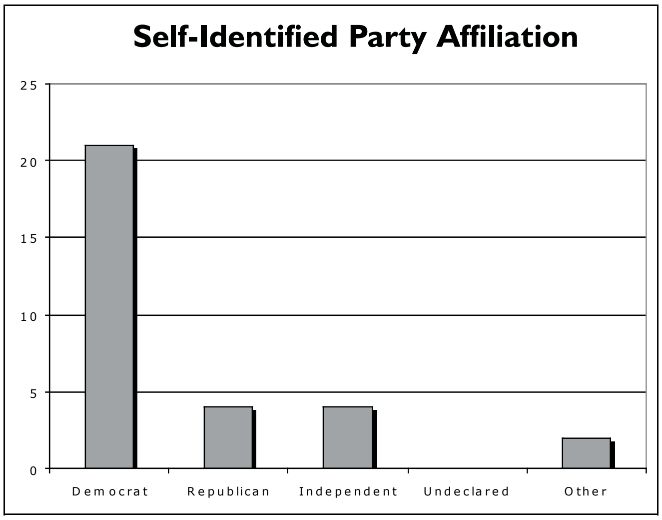


FIGURE 6. The majority of participants identified themselves as Democrats.

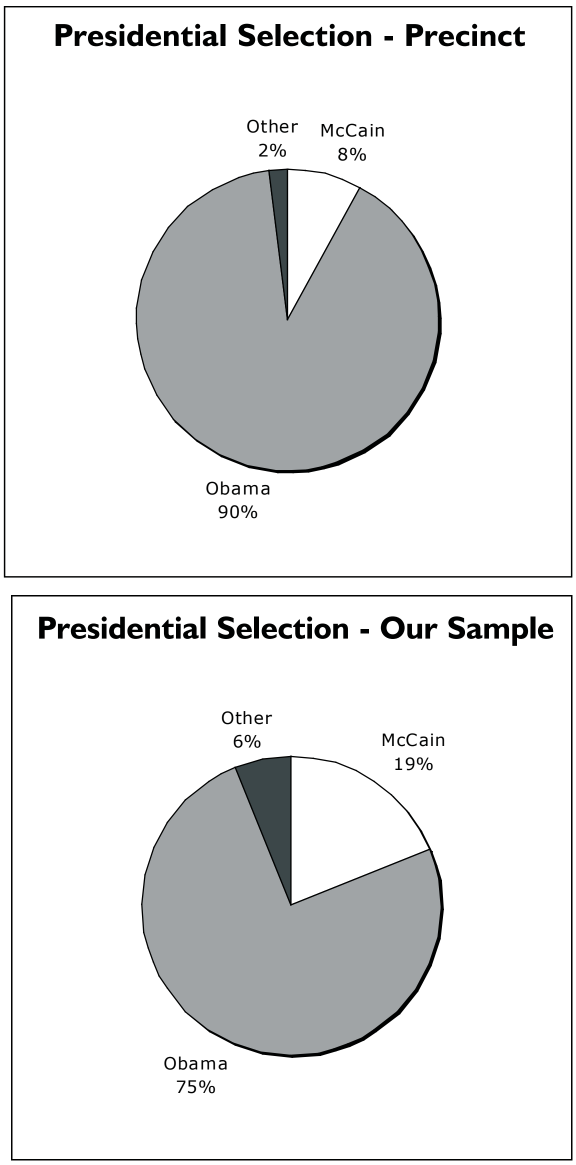


FIGURE 7. How the precinct voted compared to how the sample voted

While the following five factors were ranked as important in the voter's decision for president, they are in order here from most important to least important:

The economy
The war
Energy policy
National security
Health care

Eighty-one percent of Shoemaker and Thompson's voters decided for whom they would vote prior to a month before the election. Seventy-seven percent said having an Africa American on the ticket did not affect their decision.

None of those sampled said they "strongly approved of" or were "neutral to" the Bush administration. Everyone interviewed said they either "somewhat approved," "somewhat disapproved," or "strongly disapproved" of the current administration, with 72% saying they strongly disapproved.

Of the people polled, 87.5% said DC should have equal voting rights in the House and Senate. More than 50% said they did not know whether DC should continue the public school tuition voucher program. This made sense to Shoemaker and Thompson, given that many in the sample did not have children.

When asked for opinions regarding the vice-presidential candidate they had voted for, the McCain/Palin ticket responded with the following:

17% said they strongly approved
33% said they somewhat approved
50% said they somewhat disapproved

When asked the same question, the Obama/Biden ticket responded with the following:

50% said they strongly approved
46% said they somewhat approved
4% said they strongly disapproved

This comparison shows that Democratic voters were more supportive of their ticket's vice-presidential candidate than Republican voters were of theirs.

When asked which was a stronger influence on their selection for president, the McCain/Palin ticked responded with the following:

- 87.5% said they felt positively about their candidate
- 12.5% said they felt negatively about the other candidate(s)

When asked the same question, the Obama/Biden ticket responded with the following:

- 81% said they felt positively about their candidate
- 19% said they felt negatively about the other candidate(s)

It was assumed that Obama supporters were more positive about their candidate; however, it appears the negative feelings toward the Republican ticket were stronger than the negative feelings toward the Democratic ticket.

Data Comparison

To see how the Obama vote percentage in their survey compared to the certified results of the Obama vote percentage in the other 13 precincts in Ward 2, Shoemaker and Thompson performed a Student's t-test.

The null hypothesis of the t-test is that there is no significant difference between the average Obama percentage and the survey's Obama percentage. However, because Shoemaker and Thompson's p-value is so small (p = .0322), they could reject the null hypothesis with 95% confidence. In other words, it appears Shoemaker and Thompson's Obama percentage is significantly (statistically) lower than the average of all the other Obama percentages in Ward 2.

Conclusions

If they had it to do again, Shoemaker and Thompson would determine the busiest voting time of day and conduct their survey then. Their exit poll took about two and a half hours, and they think it would have gone faster if they would have had more voters cycling through their location. They also would have a back-up polling location in mind to avoid competition with others conducting exit polls. Finally, Shoemaker and Thompson would keep a better eye on voters exiting from the entrance to keep their count accurate. ◕

Editor's Note: *I remember checking in on the two students whose work is included above when they were polling. The light rain did not detract from the sense of optimism on that day. The polling reference used in their class was* Elections and Exit Polling, *dedicated to the late Warren Mitofsky. The underlying technology of polling that undergirds this article can be found in more detail at* www.votingsystems.us, *where the complete article appears.*

References and Additional Reading List

The references for each article in this issue of *STATS* are included in the listing below, along with suggestions for additional reading on related topics. The page number for each article is in blue.

4 A Quick Guide to Online Data Quality: Ensuring High-Quality Data from Panel Research

Balden, W. 2008. An unwanted impact. *Quirks*, January.

Bortner, B., E. Daley, and H. Lo. 2008. Is the long online panel quality nightmare over? Forrester Research, *www.forrester.com*.

Brooks, R. 2008. Internet data quality. *Quirks*, January.

Downes-Le Guin, T. 2005. Satisficing behavior in online panelists. Market Strategies International, *www.marketstrategies.com*.

Hesser, M. 2008. Each has its strengths and weaknesses. *Quirks*, January.

Knapton, K., and R. Garlick. 2007. Catch me if you can. *Quirks*, November.

Levin, A., and M.G. Holmes. 2007. Speed traps. *Quirks*, July/August.

Puleston, J., and D. Sleep. 2008. The survey killer. *Quirks*, November.

ESOMAR World Research. 2005. ESOMAR world research codes and guidelines: Conducting market and opinion research using the Internet. *www.esomar.org*.

RFL Communications, Inc. 2008. Platforms for data quality progress: The client's guide to rapid improvement of online research. *http://rflonline.com/register.aspx*.

Smith, R., and H. H. Brown. 2005. Assessing the quality of data from online panels: Moving forward with confidence. *www.hisbonline.com/pubs/HI_Quality_of_Data_White_Paper.pdf*.

10 Accept, Reject, …, or Get More Data? An Intuitive View of Sequential Tests

Bain, Lee J., and Max Englehardt. 1992. *Introduction to probability and mathematical statistics*. Pacific Grove: Duxbury Press.

Hogg, Robert V., and Allen Craig. 1994. *Introduction to mathematical statistics*. Upper Saddle River: Prentice Hall.

Mood, Alexander McFarlane, Franklin A. Graybill, and Duane C. Boes. 1974. *Introduction to the theory of statistics*. New York: McGraw-Hill.

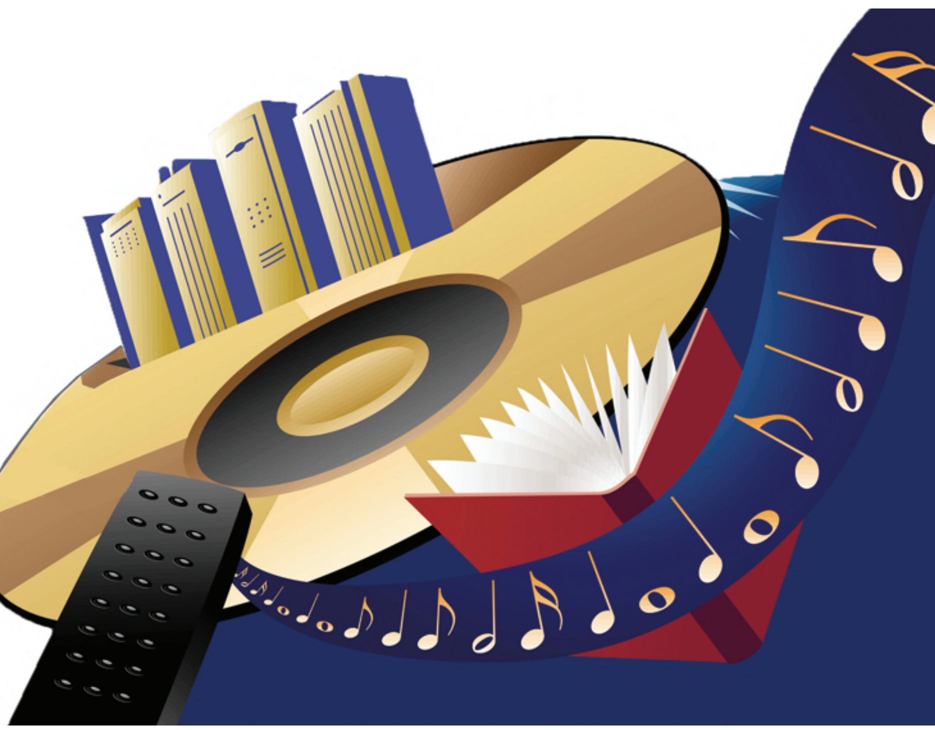
Wald, A. 1945. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics* 16:117–186.

18 Exit Polling in DC: The 2008 Presidential Election

Scheuren, Fritz J., and Wendy Alvey. 2008. *Elections and exit polling*. Hoboken: John Wiley & Sons, Inc.

Similar exit polls conducted in 2008 can be viewed at *www.votingsystems.us*.

Election day turnout for each precinct can be found at *www.dcboee.org/election_info/election_results/elec_2000/general_elec.asp*.



MAKING SENSE OF STATISTICAL STUDIES

Making Sense of Statistical Studies consists of student and teacher modules containing 15 hands-on investigations that provide students with valuable experience in designing and analyzing statistical studies. It is written for an upper middle-school or high-school audience having some background in exploratory data analysis and basic probability.

www.amstat.org/education/msss