

# STATS

THE MAGAZINE FOR STUDENTS OF STATISTICS : : ISSUE 50

**Jill Lacey and Junting Wang**  
Describe what it takes to design  
an Election Day exit poll

**Chris Olsen**  
reviews the book  
*Applied Spatial Statistics for  
Public Health Data*

## CASE STUDY:

### The *FiveThirtyEight.com*

### Predictive Model of the 2008 Presidential Election





# Looking for a **JOB?**

<http://jobs.amstat.org>

Your career as a statistician is important to the ASA and we are here to help you realize your professional goals.

The **ASA JobWeb** is a targeted job database and résumé-posting service that will help you take advantage of valuable resources and opportunities. Check out the many services available from the **ASA JobWeb**.

**VIEW ALL JOBS** → Search by keyword, job category, type of job, job level, state/country location, job posting date, and date range of job posting.

**ADVANCED SEARCH** → Use multiple search criteria for more targeted results.

**MAINTAIN A PERSONAL ACCOUNT** → Manage your job search, update your profile, and edit your résumé. (ASA members only)

**USE A PERSONAL SEARCH AGENT** → Receive email notification when new jobs match your criteria. (ASA members only)

**ADVERTISE YOUR RÉSUMÉ** → Post a confidential profile so employers can find you. Registered job seekers can submit their résumés to the résumé database in a “public” (full résumé and contact information) or “confidential” (identity and contact information withheld) capacity. A confidential submission means only an employer can contact the applicant using a “blind” email. (ASA members only)

Visit the ASA JobWeb online **TODAY**  
<http://jobs.amstat.org>

# STATS

## contents

“Sound Policies Rest on Good Information.”

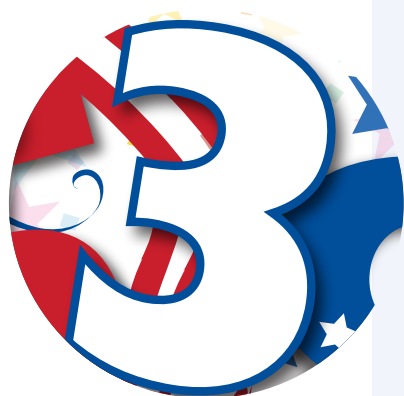
-Frederick Mosteller

## features

### SPECIAL TOPICS

CASE STUDY: The *FiveThirtyEight.com* Predictive Model of the 2008 Presidential Election

3



page 3

Designing and Implementing an Election Day Exit Poll

10

When Randomization Meets Reality An impact evaluation in the Republic of Georgia

16



page 10

## columns

### COLUMNS

EDITOR'S COLUMN 2

**RU SIMULATING**  
Matching Items and Collecting Coupons 21

**STATISTICAL U-SINGS**  
Black Death  
A review of *Applied Spatial Statistics for Public Health Data* 26

**REFERENCES** and Additional Reading List 28

### CASE STUDY: The *FiveThirtyEight.com* Predictive Model of the 2008 Presidential Election

ADAM FELDER is a consultant with *Métier, Ltd.*, and performed work on this article while a graduate student at The George Washington University.



### Designing and Implementing an Election Day Exit Poll

JILL N. LACEY is a senior survey specialist at the U.S. Government Accountability Office and performed work on this article while a graduate student at The George Washington University.



JUNTING WANG is an engineer at China Communications and Transportation Association in China.



### When Randomization Meets Reality An impact evaluation in the Republic of Georgia

MIRANDA BERISHVILI is an associate professor at Georgian Technical University in Tbilisi.



CELESTE TARRICONE is an assistant director for the Millennium Challenge Corporation, where she is a monitoring and evaluation specialist for country programs in Georgia, Honduras, and Vanuatu.



SAFAA AMER is employed by NORC at the University of Chicago and does poverty program impact evaluation in Georgia for the Millennium Challenge Corporation and its Georgian counterpart, MCG.



## EDITOR'S COLUMN

### DIRECTOR OF EDUCATION

**Martha Aliaga**  
American Statistical Association  
732 North Washington Street, Alexandria, VA 22314-1943  
martha@amstat.org

### EDITORIAL BOARD

**Peter Flanagan-Hyde**  
Mathematics Department  
Phoenix Country Day School, Paradise Valley, AZ 85253  
pflanaga@pcds.org

**Schuyler W. Huck**  
Department of Educational Psychology and Counseling  
University of Tennessee  
Knoxville, TN 37996  
shuck@utk.educoolsen@cr.k12.ia.us

**Jackie Miller**  
Department of Statistics  
The Ohio State University, Columbus, OH 43210  
jbm@stat.ohio-state.edu

**Chris Olsen**  
Department of Mathematics  
Thomas Jefferson High School, Cedar Rapids, IA 53403  
colsen@cr.k12.ia.us

**Bruce Trumbo**  
Department of Statistics  
California State University, East Bay, Hayward, CA 94542  
bruce.trumbo@csueastbay.edu

### PRODUCTION/DESIGN

**Megan Murphy**  
Communications Manager  
American Statistical Association

**Val Snider**  
Publications Coordinator  
American Statistical Association

**Melissa Muko**  
Graphic Designer/Production Coordinator  
American Statistical Association

**STATS: The Magazine for Students of Statistics** (ISSN 1053-8607) is published three times a year, by the American Statistical Association, 732 North Washington Street, Alexandria, VA 22314-1943 USA; (703) 684-1221; [www.amstat.org](http://www.amstat.org). *STATS* is published for beginning statisticians, including high school, undergraduate, and graduate students who have a special interest in statistics, and is provided to all student members of the ASA at no additional cost. Subscription rates for others: \$15.00 a year for ASA members; \$20.00 a year for nonmembers; \$25.00 for a Library subscription.

Ideas for feature articles and materials for departments should be sent to Editor Paul J. Fields at the address listed above. Material must be sent as a Microsoft Word document. Accompanying artwork will be accepted in four graphics formats only: EPS, TIFF, PDF, or JPG (minimum 300 dpi). No articles in WordPerfect will be accepted.

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

Copyright (c) 2009 American Statistical Association.

The guest editors for this issue of *STATS* are Jill Lacey and Fritz Scheuren. Lacey is a graduate of the certificate program in survey design and data analysis at The George Washington University (GWU). Scheuren, a past-president of the American Statistical Association, was her instructor, aided by Ali Mushtaq.

Some background about this change in editorship might be in order before we talk about the contents of the current issue. As many of you will remember, Paul Fields was the *STATS* editor from 2005 to 2008. He had to step down to work on other projects. We volunteered to fill in as co-editors for the time being, and we are not sorry! So far, we have found the experience stimulating and humbling—but we are not stumbling, thanks to the great staff at the ASA office, led by Megan Murphy and Martha Aliaga.

As “newbies,” we quickly learned that getting together a good set of articles is the first step to putting out an issue. Here, we were fortunate. Fields left us an inventory of fine articles from which to draw. Regular columns by Chris Olsen and Bruce Trumbo will be found in this issue, along with a topical article by Miranda Berishvili, Celeste Tarricone, and Safaa Amer about the use of randomization in agricultural experimentation. Because the experiments in the latter took place in the country of Georgia, we have provided context by reprinting an article from *Amstat News* that appeared last October.

As we all know, the United States went through a major presidential election in 2008, so we devoted the cover story to that process. Indeed, we lead with “Case Study: The *FiveThirtyEight.com* Predictive Model of the 2008 Presidential Election.” Author Adam Felder took the fall 2008 GWU certificate course on survey management. As part of the class, students conducted exit polls in the Washington, DC, area. The paper “Designing and Implementing an Election Day Exit Poll,” by Jill Lacey and Junting Wang, is an outgrowth of that course, as well.

Trumbo and Luther B. Scott wrote “Matching Items and Collecting Coupons” for the R U Simulating column. As statisticians, our love and appetite for real data in increasingly close to our love for simulated data. Trumbo and Scott tap into this growing strength.

The last article, the Statistical  $\mu$ -sings column, is a light-hearted piece from Olsen that captures the tone we wanted for this issue—interesting and instructive. This, too, comes from Fields. Thanks again!



# CASE STUDY: The *FiveThirtyEight.com* Predictive Model of the 2008 Presidential Election

by Adam Felder

**T**here has already been a great deal of analysis on the 2008 United States election in which Barack Obama was elected president. Much of this analysis is devoted to Obama's "ground game," a grassroots network of volunteers who helped register voters in record-breaking numbers and, perhaps more importantly, got those voters to actually vote on or before November 4. Much of this ground game was founded in the so-called "netroots," a term used to describe Internet-based grassroots activism.

The Obama campaign was not alone in its use of the Internet; political analysts also used it to great

effect. Large numbers of independent pollsters were able to disseminate national- and state-level polling results to the public. It is not a stretch to say that on any given business day from the time of the first presidential debate onward, at least a dozen polls from various organizations were released.

To combat this saturation of information, poll aggregation sites exist. These sites compile all polling for a given area (e.g., a state or national poll) and combine it in an attempt to give a clearer picture of the state of the race. However, there exist different methodologies for aggregation. Some lump all polls together and treat them equally. Others weight by timeliness of a poll; a poll taken a week before Election Day is given greater weight than a poll taken a month before Election Day. Still others, such as *FiveThirtyEight.com*, allow polls from demographically similar states to influence one another's results in the absence of new information. In short, various aggregation sites employ mixed methodologies to make predictions.





## About *FiveThirtyEight.com*

*FiveThirtyEight.com*, named for the number of electoral votes awarded in a presidential election, is owned and operated by Nate Silver. The site launched in March 2008, and Silver (using the pseudonym “Poblano”) posted state-by-state predictions for the still-ongoing Democratic primary battle between Barack Obama and Hillary Clinton. Prior to March, Silver, posting on the web site Daily Kos ([www.dailykos.com](http://www.dailykos.com)), predicted the outcomes of the Super Tuesday primaries with a great deal of accuracy, giving his future predictions legitimacy.

In late May 2008, Silver gave up his pseudonym and revealed his identity. For those who follow baseball, this gave Silver even more legitimacy in his political predictions. Silver is a writer for *Baseball Prospectus*, a web site focusing on Sabermetrics.

Sabermetrics is the term used to describe statistical studies pioneered and performed by the Society for American Baseball Research (SABR), which has found numerous statistical indicators of player performance and value to a team that are not printed on baseball cards or in newspaper box scores. Sabermetric analyses give a better summary of player performance and a team’s long-term winning potential. Indeed, Silver was one of just a few to correctly predict that the Tampa Bay Rays, perennially a last-place baseball club, would enjoy success in 2008.

With his baseball pedigree lending credence to his political predictions, Silver gained a lot of traffic to his site. One of the nicer features of the site for observers is that Silver frequently posts updates to his methodology; there is a great deal of transparency in his analysis. This allows onlookers to examine the predictive success of the *FiveThirtyEight.com* model.

Because the final results were not close, none of the various methodologies employed resulted in a prediction different from the actual result—an Obama victory. Had the pre-election polling been closer, it seems likely that some sites would have been on the wrong side of history in their predictions. I’m going to examine the methodology of one such site, *FiveThirtyEight.com*, and analyze what aspects of the methodology were ultimately most predictive. It is important to remember that the 2008 election is over and we are not looking to design a predictive model that perfectly fits the 2008 results, but rather a model that is flexible enough to be applied successfully to previous and future elections.

## Methodology

*FiveThirtyEight.com* has several unique features to its model that differentiate it from similar sites. The first major difference is in how polls are weighted. Additionally, the model includes a regression estimate that helps reduce the impact of outlier results. A trendline is established from the results of national polling, which is applied toward states that are polled infrequently. As mentioned previously, states that are demographically similar are grouped so a poll in one state has influence on similar states. Finally, the model runs an election simulation 10,000 times daily to account for outlier predictions and to provide the most likely outcomes.

## Weighting

In the *FiveThirtyEight.com* model, weighting is not merely a function of the timeliness of a poll. It also takes into account the sample size of a poll and the historical accuracy of the organization conducting the poll.

The timeliness portion is fairly intuitive. To give an extreme example, do not give equal weight to two polls, one conducted the day before the election and one conducted two months prior to the election. The freshness of a poll is important in assessing its reliability because attitudes toward a candidate can change. For *FiveThirtyEight.com*, the weight assigned to a poll can be expressed as  $0.5^{(P/30)}$ , where P is the number of days transpired since the median date the poll was in the field.

Weighting by sample size takes into account the margin of error introduced by a given sample. It is expressed by the formula  $80 * n^{(-.5)}$ , where n is the number of respondents in the poll. Given variance, the margin of error will be greater when the actual result is close to an even distribution and lesser in the case of large landslides.

The accuracy of a pollster (how close a pollster is to the true mean) also is included in weighting, but should not be confused with the reliability of a pollster (how close a pollster’s numbers are to one another in repeated results). A pollster who consistently overstates the support for a candidate by two points can be adjusted to represent the true level of support. A pollster whose results are consistently inconsistent is much more difficult to fit into a statistical model. The *FiveThirtyEight.com* model refers to this inconsistency as “pollster-introduced error,” or PIE. Put simply, polling outfits with a lesser amount of PIE are weighted more heavily than those with greater amounts of PIE.

## Trendline Adjustment

Even with the amount of polling seen in 2008, polls tended to focus on swing states. States where the outcome was known well before candidates were nominated from either party (e.g., Connecticut was a guaranteed Democratic pickup, while Utah was a guaranteed Republican win), especially those with small electoral vote totals, did not seem to be particularly interesting to pollsters and, as such, were polled infrequently. *FiveThirtyEight.com*'s trendline seeks to compensate for this lack of polling by applying known results to these unpolled states.

The model divides data points (individual polls) into state-pollster units and week units (incrementing by one per week approaching November 4). Additionally, the national numbers are classified as their own state. With these data points, a regression line is applied and employs a Loess smoother. This allows a user to see any trend in recent polling numbers and apply it to current polling.

What this ultimately means is that if a candidate's numbers have generally improved 10 points from where they were on a given date, it can be reasonably inferred that the candidate's numbers have improved by a similar amount in a state that has not been polled since that date. Granted, this is an oversimplification; demographic trends in individual states also exert influence on polling numbers. However, it would be similarly incorrect to assume no change occurred in a recently unpolled state. The *FiveThirtyEight.com* model takes these demographic differences into account to try to mitigate against this type of error.

One of the weaknesses of this adjustment is that it will be overly sensitive to trends, serving as a snapshot of support, rather than a prediction of support on Election Day. To use a concrete example, Obama's support surged across many demographic groups following the Democratic National Convention (DNC). With the trendline indicating this surge, the *FiveThirtyEight.com* model indicated a massive Obama blowout. When John McCain experienced his own surge following the Republican National Convention (RNC), the trendline captured this as well, predicting a convincing victory for McCain.

In both instances, the trendline was oversensitive to "bounces," temporary spikes in one candidate's support. Indeed, while Obama did win convincingly, both in the electoral and popular votes, another bounce was a major factor in his margin of victory: the public collapse of Lehman Brothers and the subsequent focus on the struggling American economy. Obama's numbers improved across the board from that date forward.

Chart 1 is courtesy of another aggregation site,



CHART 1. Bounces in late August (the DNC), early September (the RNC), and the ongoing Obama bounce stemming from the problems on Wall Street

*Pollster.com*. This site offers users the ability to create custom graphs for selected date ranges and polling outfits, a feature not available on *FiveThirtyEight.com*. In the chart, you can see "bounces" in late August (the DNC), early September (the RNC), and the ongoing Obama bounce stemming from the problems on Wall Street.

This shows another weakness in *FiveThirtyEight.com*'s operation. In early August, Silver detailed a change to his model that would attempt to compensate for predictive "convention bounces" for both candidates. This change would make the model temporarily less sensitive to trendline adjustments for a length of time determined by the average length of a convention bounce for candidates in previous elections. However, as post-convention polls came out showing Obama with a sometimes double-digit lead, pressure from the *FiveThirtyEight.com* user community prompted Silver to post a user poll on August 31.

This poll would determine whether Silver would remove the convention bounce adjustment from the model. Given that the demographics of visitors to the site trended overwhelmingly Democratic, voters expressed overwhelming support for the removal of the adjustment. Thus, with no statistical reason to do so, the *FiveThirtyEight.com* model was modified to show a massive shift toward Obama, and later a massive shift toward McCain, all in the span of approximately two weeks.

During this period, the model was highly volatile and likely not predictive of anything beyond an election that would have occurred on the same day as the prediction, rather than any results for November 4.



## Regression Estimate and Demographic Similarity

National elections generally address issues that are applicable to the entire country, rather than just a single state. That said, many issues tend to be viewed in different ways by different demographic groups. To use a recent example, Obama received the overwhelming share of African-American votes nationally, winning, on average, more than nine out of 10 voters. It is logical to conclude, therefore, that this particular demographic would be a benefit to Obama in states where there is a large African-American population. This does not mean, however, that all states with large African-American populations would end up in Obama's column; there are clearly many other demographic factors at play.

Indeed, polling results in Mississippi, for example, would indicate exactly that (McCain: 57%, Obama: 43%, despite 37% of the population being African-American). *FiveThirtyEight.com*'s model attempts to group states into demographic categories to predict results in states that were polled less frequently. The variables considered in determining the demographic makeup of each state were classified into several subcategories.

The "political" subcategory encompasses four variables that generally measure a state's overall political leaning, its support for each major candidate in the 2004 election, and the fundraising share for either of the two major candidates. To compensate for "home state advantage" (i.e., Massachusetts and Texas) when examining 2004 support, the most recent candidate from that party not from that state is used. For example, Al Gore's 2000 Massachusetts support is used in lieu of Kerry's 2004 counterpart. While Obama's 2004 DNC speech claimed, "There are no red states and blue states," the model attempts to lump states into exactly those categories, or at the very least, on a point in between.

The "religious identity" subcategory measures the proportion of white evangelical Protestants, Catholics, and Mormons in each state. Historically, all three groups trend toward supporting the Republican candidate, though exit polls on November 4, 2008, showed a difference in this trend. While Mormons and Protestants supported McCain, Obama outperformed McCain among Catholic voters in exit polls by a margin of 54% to 45%.

The "ethnic and racial identity" subcategory measures the proportion of African Americans, Latinos, and those self-identifying as "American" in each state. In the case of the Latino population, this is measured by voter turnout in 2004 to compensate for new migrants who are not yet citizens. The "American" variable tends to be

highest in the Appalachian areas of the country, areas in which political pundits from the time of the Democratic primaries onward predicted Obama would struggle.

Economic variables encompass per capita income by state, as well as the proportion of jobs in the manufacturing sector.

Demographic variables cover specific ages—the proportion of residents ages 18–29 and, separately, the white population 65 or older. These two demographic groups trend toward the Democratic and Republican candidates, respectively, and it was hypothesized that the leaning would be even stronger given the Obama vs. McCain match-up—a finding that was confirmed by national exit polling on November 4, 2008. Education level and suburban residency rounded out the demographic variables.

Ultimately, all these variables are fairly intuitive and the sort you might expect to see in an exit poll. While Silver would occasionally experiment with 'fun' new variables to see if they were significant (at the 85% level) indicators of candidate support, the aforementioned variables tended to be the best indicators.

Using these indicators allowed the *FiveThirtyEight.com* model project results for states that were under-polled. For example, Kentucky was rarely polled, but West Virginia was polled fairly frequently. West Virginia and Kentucky were found to be demographically similar, and thus West Virginia's polling numbers exerted some influence on those in Kentucky.

## Simulation and Projection

With all these factors in mind, the *FiveThirtyEight.com* model ran a simulated election 10,000 times nightly. With this sample size, the user could see what outcomes were most likely to occur and what the mean outcome was. Perhaps more interestingly, as Obama pulled further away from McCain in the final month, a user could study the McCain victory scenarios to see what states were most critical. McCain's strategy for winning Pennsylvania, while ultimately unsuccessful, is somewhat understandable when viewing the few simulations in which McCain received at least 270 electoral votes.

## Outcome

*FiveThirtyEight.com*'s final model on the morning of November 4, 2008, predicted a 98.9% chance of an Obama victory—with Obama receiving 52.3% of the popular vote to McCain's 46.2%—and a final electoral vote score of 349 to 189. The prediction turned out to miss the popular vote difference by 0.6%. The model incorrectly predicted only Indiana and the 2nd Congressional District of



Nebraska, both of which Obama won. This seems a strong performance.

## Analysis

Even with a highly sophisticated model, a polling aggregation site such as *FiveThirtyEight.com* is limited by the quality of the polls composing the aggregation. As in any population measure, a small sample size is more likely to be influenced by outliers. When this is applied to state-level polling, states where fewer polls were conducted are less likely to accurately capture the true mean of support for either candidate.

Pollsters have limited resources; it is to their benefit to deploy those resources in states where the outcome is somewhat in question. Thus, it is not terribly surprising to find that the 10 most-pollled states (shown in Table 1) were swing states (or in the case of Pennsylvania and Minnesota, very publicly targeted by the McCain campaign as winnable blue states).

State	Number of Polls Within Six Weeks of Election	Final Margin of Difference	Winner
OH	43	4	Obama
FL	38	2.5	Obama
PA	35	10.4	Obama
NC	34	1.1	Obama
VA	32	5.5	Obama
MO	25	0.2	McCain
CO	23	8.6	Obama
MN	21	10.2	Obama
IN	20	0.9	Obama
NV	20	12.4	Obama
	Average Margin	5.58	Obama

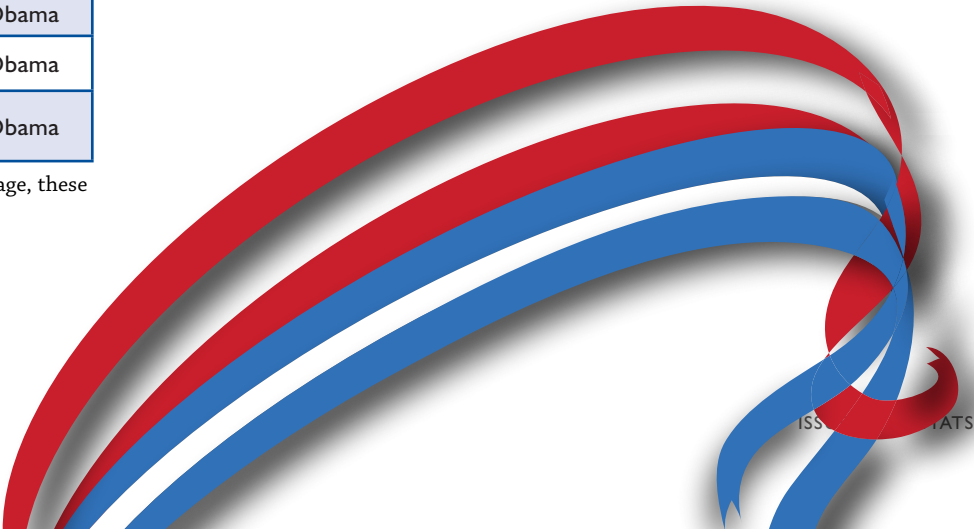
TABLE 1. The 10 most-pollled states. On average, these states were decided by 5.6 percentage points.

State	Number of Polls Within Six Weeks of Election	Final Margin of Difference	Winner
DC	1	86.4	Obama
RI	2	27.8	Obama
NE	2	16.1	McCain 4 Obama 1
MD	2	24.1	Obama
ID	2	25.4	McCain
HI	2	45.2	Obama
CO	23	8.6	Obama
UT	3	28.7	McCain
VT	4	35.2	Obama
SD	4	8.5	McCain
ND	4	8.6	McCain
	Average Margin	30.6	

TABLE 2. The 10 least-pollled states. On average, these states were decided by 30.6 percentage points.

Meanwhile, the 10 least-pollled states (shown in Table 2) were decided by an average of 30.6 percentage points. It is worth pointing out that while McCain won Nebraska, its apportionment of electoral votes is done by congressional district. Despite losing the state, Obama won the 2nd Congressional District. In general, however, pollsters tended to ignore states that were blowouts. Thus, while *FiveThirtyEight.com*'s model may have been susceptible to outliers, these outliers would only have served to throw off the margin of blowout, rather than change any predictions of the actual winner.

Note: Polling data were taken directly from *FiveThirtyEight.com* tables. Final reported margins by state were taken from *Pollster.com*.



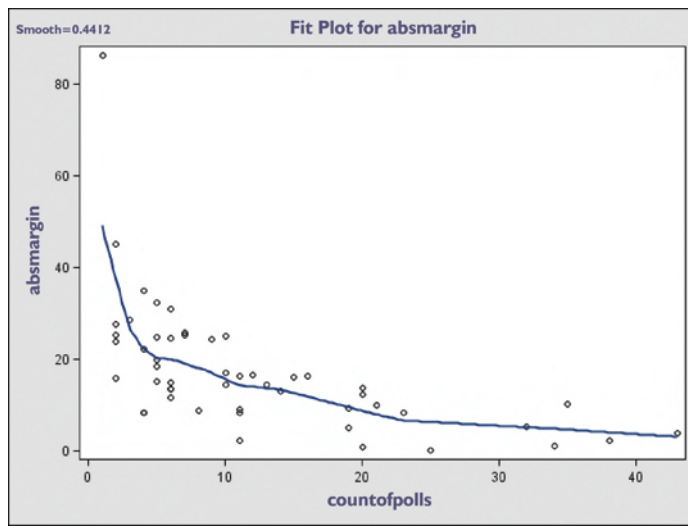


CHART 2. The number of polls used by *FiveThirtyEight.com*'s model for a given state versus its reported margin of difference between the candidates

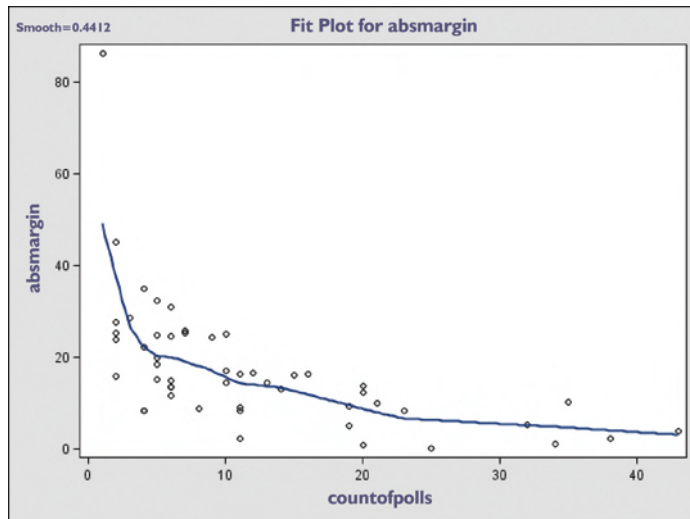


CHART 3. The number of polls in a state plotted against the difference between the *FiveThirtyEight.com* prediction and the reported results of the 2008 election

Chart 2 shows the number of polls used by *FiveThirtyEight.com*'s model for a given state versus its reported margin of difference between the candidates. Pollsters generally targeted the correct states in the 2008 election; there were noticeably more polls taken in tighter states.

Furthermore, the number of polls focused in so-called "swing" states proved beneficial to *FiveThirtyEight.com*'s model. The more data it had to work with, the lesser the margin of difference between its predictions and the reported results of the election.

Chart 3 displays the number of polls in a state plotted against the difference between the *FiveThirtyEight.com* prediction and the reported results of the 2008 election. In general, two conclusions can be drawn from this chart:

The more polling available in a state, the more accurate the *FiveThirtyEight.com* model becomes

The accuracy of results experience diminishing returns as the number of conducted polls increases

A great deal of accuracy is gained for the first five to six polls of a state, with lesser amounts of accuracy gained from the seventh to eleventh polls. From that point forward, there is a steady (though small) return on investment.

None of these findings is groundbreaking, but they do confirm that pollsters should really focus their efforts in states they believe will be determined by fewer than four points.

As stated previously, the 2008 presidential election was ultimately not close. While the *FiveThirtyEight.com* model took a great number of external variables into account, an unsophisticated aggregation would have returned similar results.

Removing controls for weight (a function of poll date, sample size, and polling organization), as well as the trendline function, you can run a simple aggregation.

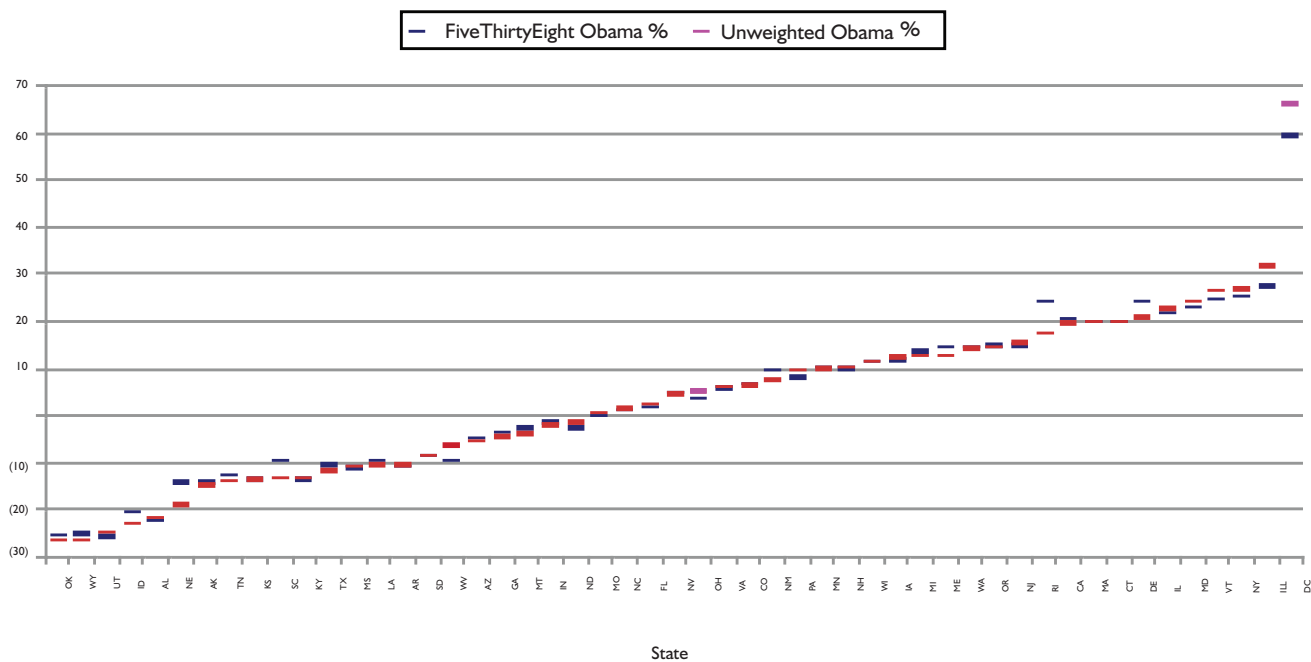
This aggregation, which we'll call "unweighted aggregation," calculates the mean support by state for both Obama and McCain by adding their support across all polls and then dividing by the number of polls. One minor adjustment is made to account for undecided voters. Undecided voters—expressed as "100-(McCain support+Obama support)"—are allocated 50-50 to each of the two candidates. While this excludes the potential for third-party candidate support, the impact of this neglect is minimal and outweighed by the benefit gained from bringing the unweighted aggregation's prediction closer in line with the reported final results.

Using this aggregation, only one state's prediction differs from the *FiveThirtyEight.com* prediction: Missouri. Missouri, which McCain ultimately won by less than a point, was predicted to be a two-tenths victory for McCain by the *FiveThirtyEight.com* model. The unweighted prediction saw Missouri as a six-tenths victory for Obama. Indiana was called incorrectly in both models by similar margins. By and large, the tweaks in the *FiveThirtyEight.com* model did not have a great deal of influence on its predictions; a comparison of the two predictive models by state shows nearly a perfect overlap, as shown in Chart 4.

There is a correlation (significant at the 95% level) between the number of polls taken in a state and the difference (either positive or negative) between the reported results and the prediction of the *FiveThirtyEight.com* model.

Many of the states with few polls had their polls conducted by a nonprolific pollster such as Rasmussen, SurveyUSA, or YouGov. (For the purposes of this analysis, "prolific" will be defined as the top three agencies by number of polls included in the model.) Thus, removing these

## Predicted Margin vs. Reported Margin



**CHART 4.** A comparison of the reported margin of victory for Obama (negative numbers indicate McCain won that state) vs. the predicted margins of victory by the *FiveThirtyEight.com* model and unweighted aggregation

minor agencies from the model would not give a complete picture of a pollster’s impact in all states across the country. Instead, we will examine the unweighted model’s prediction if one of these three pollsters did not have its polls included. The *FiveThirtyEight.com* prediction cannot be re-run in the same manner, but given how closely aligned the *FiveThirtyEight.com* and unweighted models are, one can reasonably assume the same results.

Perhaps not surprisingly, there is little shift in the predictive model when any prolific agency is excluded from the model. The unweighted model continues to predict Missouri and Indiana incorrectly. This, too, is rather unsurprising. As previous analysis indicates, the more polls used by the *FiveThirtyEight.com* model, the more accurate its prediction. Additionally, the unweighted model is virtually identical to the *FiveThirtyEight.com* model. Furthermore, prolific agencies are concentrated in states whose margin of difference is small between the two candidates; there is little polling of blowout states by such agencies.

Thus, when removing one of these prolific agencies from the model, several data points may be removed from a data-heavy state, but few, if any, data points are removed from states whose outcome was known before polls opened on November 4.

It seems a logical conclusion that, in 2008, enough polling existed to account for any error made by a single pollster—no matter how prolific. Tiny pollsters, even if they were in error, did not

have enough proliferation to damage the prediction, and large pollsters had their data points mixed in with their counterparts. The 2008 election is one in which the pollsters were generally spot-on; the *FiveThirtyEight.com* and unweighted models benefited tremendously from this fact.

### Conclusion

The *FiveThirtyEight.com* model is seemingly useful, but not significantly more useful than a less-sophisticated model. The methodology behind its predictions are a supplement, but not a replacement for, actual data. Furthermore, in states where there is a great deal of polling, the availability of data appears to overwhelm other elements of the methodology. That said, pollster resources are finite, not all elections will be as uncompetitive as the 2008 election, and the *FiveThirtyEight.com* model did get one more state correct than the unweighted model. In closer elections, the *FiveThirtyEight.com* model could make the difference in reporting which candidate wins. Regardless, the ingenuity shown by Silver in capturing trends in a predictive model, as well as applying results from demographically similar states to one another, is an interesting and noteworthy study in management. ●

---

Editor’s Note: *In 2008 we lived through yet another election; this one was much less controversial than those of 2000 or 2004. Politics aside, as a statistical matter, 2008 was more satisfying in that we ended up with a widely accepted prediction that turned out to be quite good.*

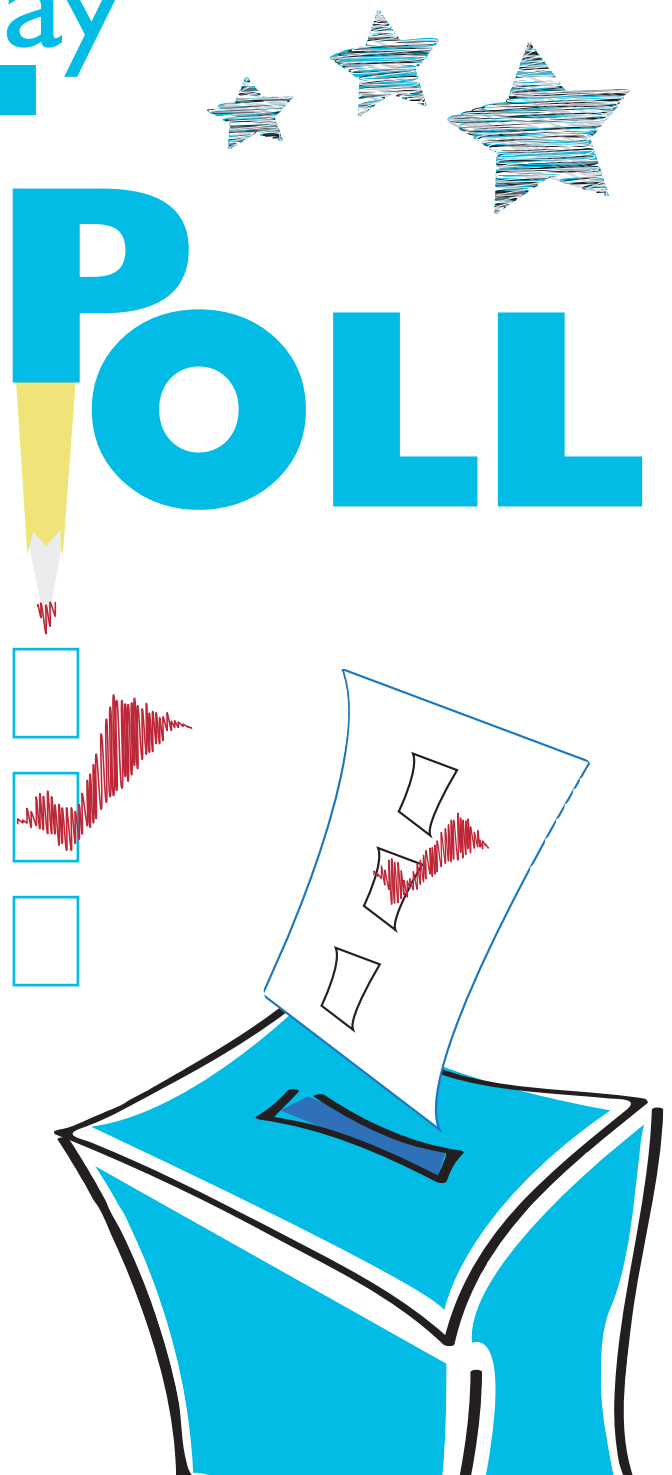
# Designing and Implementing an Election Day EXIT POLL

by Jill Lacey and Junting Wang

**T**he fall of 2008 was an exciting time to be a student of survey management at The George Washington University. To gain hands-on experience managing a small survey, we devoted the majority of the semester to designing, implementing, and analyzing results from our own Election Day exit poll. Students worked in teams of two and chose precinct locations throughout the Washington, DC, metropolitan area. Four groups, including ours, chose to conduct their exit polls at different precincts in Alexandria, Virginia. While the four groups worked together to develop a questionnaire, each pair was responsible for their own data collection and analysis.

## Questionnaire

Designing the questionnaire proved to be one of the most difficult tasks in the project. The four Alexandria groups worked together on the questionnaire to properly pre-test the questions. Having one questionnaire also increased the number of observations available to other researchers who might want to combine data from the four precincts.



Our objective was to find out which candidate respondents voted for and why. This helped us narrow our choice of questions to fit on one page. We also asked basic demographic questions, such as age and race, to aid in the nonresponse and data analysis.

Each polling group had four versions of the questionnaire. Two sets of questionnaires alternated Barack Obama and John McCain as the first candidate listed in the response categories for the presidential horse-race question. Although respondents were not aware of this randomization because they only received one copy of the questionnaire, this reduced the potential bias of the questionnaire designers, who might have unknowingly listed their favorite candidate first.

Another set of questionnaires indicated which group member intercepted the respondent. One group member used questionnaires that had section headers highlighted, and the other member had questionnaires with section headers underlined. Again, this subtle difference went unnoticed by respondents. The use of different indicators allowed us to analyze how interceptor characteristics influenced who was likely to cooperate and how those respondents voted.

The questionnaires were also translated into Spanish because of the growing Hispanic population in Northern Virginia. However, only one respondent in the four precincts used the Spanish-language questionnaire.

The questionnaire went through cognitive testing and pre-testing at the early voter precincts before Election Day. Based on results of the tests, questions were appropriately modified. For example, terrorism was dropped as one of the issues important to voters and replaced by energy policy, which became an important issue during the 2008 campaign because of high gasoline prices in the months preceding the election.

However, an error remained on one set of the questionnaires that went undetected until the data were already collected. In the questionnaires with the underlined sections, which indicated one of the interceptors, the important issues question included a fifth category of “other,” whereas the questionnaires with the highlighted section headers only listed four categories—the economy/taxes, foreign policy, health care, and energy policy. During analysis, the answers to the “other” option were dropped from the half of respondents who received this questionnaire.

## Precinct

The four Alexandria groups randomly selected their precincts from a total of 24 precincts using

a table of random numbers. Polling precincts are usually chosen using probability proportional to size sampling methods to give the most populous precincts a higher probability of being sampled. However, the goal of our exit polling project was not to precisely measure the vote outcome. The goal was to learn how to manage the steps in the process, so we felt it was sufficient to randomly select precincts.

Our randomly chosen precinct was an elementary school. Based on our observations from driving around the precinct, it appeared to be one of the wealthier precincts in Alexandria. We noticed about an equal number of Obama and McCain signs in yards and car windows. We learned from talking with voters and campaign volunteers at the precinct that it used to be a solidly Republican precinct, but has been trending more Democratic in recent years. In the 2004 presidential election, 63% of voters in this precinct voted for the Democratic candidate, John Kerry, and 36% voted for the Republican candidate, George W. Bush.

## Polling and Voting Conditions on Election Day 2008

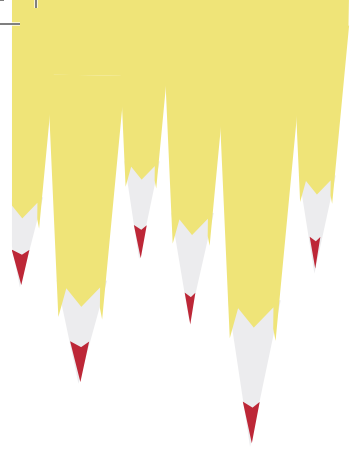
Throughout the duration of our polling, we observed factors that could affect poll results, such as weather and voter traffic. We conducted our poll during the middle of the day. We began conducting the exit poll at about 2:30 p.m. and ended around 4:30 p.m. It rained throughout the afternoon at varying intensities, from a slight drizzle to a steady downpour.

Virginia has a 40-foot, no-campaigning rule around all poll entrances. The school only had one entrance and one exit, and the exit was at the opposite end of the building from the entrance. The voting took place in the school's gymnasium, and voters exited from the gymnasium directly onto the sidewalk outside. Because the exit was more than 40 feet from the entrance, we were able to stand very close to the door to intercept voters.

There were no other exit pollsters at our precinct when we were there. There were campaigners for both Obama and McCain, as well as several members of a vote monitoring group. These other groups did not impede us in any way from intercepting voters and were generally cordial and interested in our project.

When we arrived at the polling place, we immediately introduced ourselves to the election officials inside and informed them that we would be conducting an exit poll for a couple of hours. The election officials were cooperative and did not express concern about our presence.

When we introduced ourselves to the election officials, we noticed there was no line of people



waiting to vote. This was probably because of the time of day we conducted our poll. By 2:30 p.m., about 83% of voters in this precinct had cast their ballots. Election officials told us 1,927 people had already voted. A total of 2,323 people in the precinct voted on Election Day. When we concluded our polling at 4:30 p.m., the total number of people who voted was 2,191; therefore, only 264 voters came to the polls during the two hours we were there.

### **Polling Protocol**

The two requirements we had for the exit poll was that we had to have a minimum sample size of 30 and both group members had to act as an interceptor and recorder.

### **Interceptor and recorder tasks**

The interceptor's job was to approach voters as they left the polls and try to get them to complete the survey. The recorder's job was to record the outcome of each attempt (completed interview, refusal, miss, or not eligible) and the apparent sex, age, and race of the respondent. This latter information could be used later to analyze and adjust for nonresponse.

Before Election Day, we pre-tested our protocol outside the early voter polling precinct in Alexandria. During our pre-test, we determined the ideal way to divide the interceptor and recorder jobs was to trade jobs after 10 attempts, regardless of the outcome of the attempt. This allowed us to "take a break" from the interceptor role so we would not become too fatigued. Each interceptor achieved at least 15 completes.

When we approached a potential respondent, we first asked if they had just voted so we could screen out people who were not included in the population. Examples of nonvoters who left the polling place included poll workers, mail carriers, and people accompanying voters. If a person indicated they had voted, we explained that we were The George Washington University graduate students conducting an exit poll as a class project. If the potential respondent expressed reluctance, we explained it was only a one-page survey and would take 2–3 minutes to complete. While it was raining, we also offered to hold large umbrellas over the respondent.

Each of us had two clipboards with our unique questionnaires attached. After several missed respondents at the beginning of the polling, we decided that having multiple clipboards could significantly reduce the number of misses we had. Respondents filled out the questionnaires and then

placed them completed in a covered ballot box to maintain respondent confidentiality.

During the first hour of our exit poll, our professor and teaching assistant—Fritz Scheuren and Ali Mushtaq, respectively—monitored our protocol to ensure we made every attempt to collect high-quality data.

### **Sampling interval**

Because we conducted our poll during the middle of the day when voter traffic was slow, we used a low sampling interval to get the required 30 completes in a reasonable amount of time. We used a 1-in-4 sampling ratio throughout the duration of our poll, which allowed us to collect our sample in about two hours. We tried not to deviate from the sampling interval because this can introduce significant interviewer selection bias. If two potential respondents exited the polling place at the same time, the first person who crossed a predetermined crack in the sidewalk was approached.

This interval resulted in few missed respondents. We experienced most of the misses at the beginning of the poll, but this was mainly due to two large groups of people leaving the polls at the same time and not enough clipboards with surveys for each respondent to complete the survey.

We had one voter approach us and volunteer to complete the survey, even though this individual was not part of the sampling interval. We allowed this individual to complete the survey, but put a mark on it so we could later identify it. This case was not included in the analysis.

### **Nonresponse Analysis**

It is always important to analyze nonresponse because if it is not random, it can introduce bias into the results. Interceptor bias can occur if one interceptor's characteristics make it more or less likely that he or she obtains refusals. Respondent bias can occur if respondents in certain demographic groups are less likely to respond to the poll.

### **Nonresponse by interceptor**

Our overall response rate was 68%, and our cooperation rate was 79% (see Table 1). Our response rates were higher than rates usually achieved in exit polls. Our cooperation rate was much higher than our response rate because we had several missed respondents at the beginning of our exit poll while we were setting up. However, throughout the remainder of the exit poll, each interceptor had two clipboards with questionnaires in case voter traffic leaving the polling place was heavier than normal. We think having multiple clipboards significantly reduced our misses throughout most of the poll. We

	RESPONSE OUTCOME					RATES	
	Completes	Refusals	Misses	Not Eligibles	Total	Response	Cooperation
Total	30	8	6	1	45	0.68	0.79
Jill	15	3	5	1	24	0.65	0.83
Junting	15	5	1	0	21	0.71	0.75

TABLE 1. Response outcome by interceptor

	RESPONSE OUTCOME					RATES	
	Completes	Refusals	Misses	Not Eligibles	Total	Response	Cooperation
Sex							
Male	13	6	4	1	24	0.57	0.68
Female	17	2	2	0	21	0.81	0.89
Age							
18-34	7	1	3	0	11	0.64	0.88
35-54	14	5	3	1	23	0.64	0.74
55 and over	9	2	0	0	11	0.82	0.82
Race							
Black	6	1	2	0	9	0.67	0.86
White	21	7	3	1	32	0.68	0.75
Other/ Don't know	3	0	1	0	4	0.75	1.00

TABLE 2. Response outcome by voter demographic

also credit being able to stand close to the exit for our low number of misses (six in total); almost every voter had to walk past us to get to the parking lot.

Given the rainy weather, we also had a low number of total refusals (eight in total). We believe telling voters the survey would only take 2–3 minutes helped persuade many to complete the survey. Also, offering to hold umbrellas over respondents while it was raining helped.

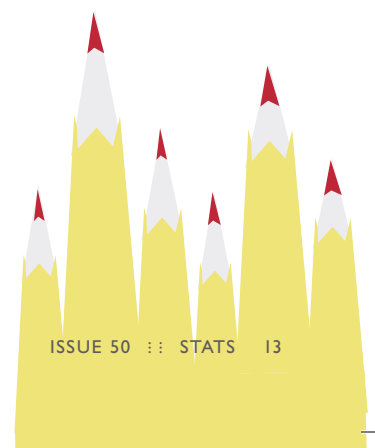
The response and cooperation rates did not differ significantly between interceptors in an independent *t*-test (response rate *p*-value=0.44; cooperation rate *p*-value=0.89). It is important to compare response rates by interceptor because interceptor characteristics can influence voters' likelihood of responding to exit polls. Both

interceptors are female and in the 18–34 age category. One is white, and one is Asian. Even though one is not a native English speaker, this did not appear to affect cooperation rates.

### Nonresponse by voter demographic

Table 2 provides a breakdown of response outcomes by voter demographic. Response and cooperation rates across different race categories did not vary greatly. The response rate for voters age 55 and over was about 18 percentage points higher than in younger age categories, but this is mainly due to misses in the younger age categories. Cooperation rates for the different age groups were similar.

Both response and cooperation rates by sex differed greatly. Women had an almost 20



NUMBER OF OBSERVATIONS		
	Guess	Actual
<b>Sex</b>		
Male	13	13
Female	17	17
<b>Age</b>		
18-34	7	11
35-54	14	9
55 and over	9	10
<b>Race/ ethnicity</b>		
Black	6	6
Hispanic	1	2
White	21	21
Other/ don't know	2	1

TABLE 3. Observed vs. actual respondent demographics

percentage point higher cooperation rate than did men. However, an independent *t*-test on these groups revealed that the difference was not significant ( $p$ -value=0.12). Overall, we were successful in persuading a demographically diverse group of voters to participate in our exit poll.

Among the voters who refused, six were white males and most were in the older age categories. This is not unexpected, because men are less likely

in general to participate in exit polls. Since both of us are young, our age may have affected our refusals among older voters. This would not be inconsistent with other exit poll results that show older voters were less likely to cooperate if the interceptor was young.

The nonresponse analysis is based on our observations of the voters' sexes, ages, and races. It is entirely possible that our observations were incorrect and that the nonresponse analysis does not reflect the true characteristics of the voters. However, we can compare our "guesstimates" for respondents with their actual self-reported demographic data from their exit poll questionnaires. We cannot match the observed and actual values for individual respondents, but we can look at the overall distribution.

We guessed respondents' sexes with 100% accuracy (see Table 3). We were almost 100% accurate guessing respondents' races.

For the three age groups, we tended to group people in the middle age category. It can be hard to determine people's ages, especially those who appear to fall right on the dividing line between age groups. Just to be sure our guesses were not statistically different from the actual ages, we did independent *t*-tests to confirm there is no difference between the two distributions (35-54 to 18-34,  $p$ -value=0.63; 35-54 to 55+,  $p$ -value=0.39).

It appears we were accurate in guessing voters' demographics and can have confidence that our observed values for the nonrespondents are also accurate.

### A Note On Statistical Weights

In our poll, men had a much lower cooperation rate than did women. Even though the rates were not statistically different, there is still a chance that nonresponse bias could exist in the estimates. Weights were calculated as  $[\text{Respondents} + \text{Nonrespondents}] / \text{Respondents}$  for both men and women. The weight was higher for men (1.85) than for women (1.24) because men had a higher nonresponse rate. When calculating the unweighted percent of the vote for Obama, where a 1 = vote for Obama and a 0 = vote for McCain, men's and women's votes count equally and sum to the number of respondents (30). With the weighted percent of the vote, men's votes count for 1.85 and women's count for 1.24. The sum of the men's and women's weights is 45, which is the total number of people sampled. In order to account for nonresponse among men and women, a weight was applied to estimates that accounted for higher nonresponse among men. However, the unweighted and weighted vote percentages for Obama and McCain were virtually unchanged. The unweighted Obama percentage of the vote was 76.67 percent, and the weighted percent was 76.71 percent.

### Polling Results

This analysis presents unweighted results because a simple nonresponse adjustment for sex resulted in virtually no change in the estimates. Additional analysis can be found at [www.votingsystems.us](http://www.votingsystems.us).

The findings of our exit poll showed Obama won the presidential horse-race question with 76.7% of the vote (see Chart 1). This was almost 14 percentage points higher than the actual vote outcome for this precinct. However, an independent *t*-test indicated this difference was not significant ( $p$ -value=0.58).

The findings of our exit poll were much closer to the actual election results for the entire city of Alexandria, which voted for Obama by a larger margin than our precinct. Our poll only differed by 0.5 percent points from the actual Alexandria vote outcome. Even though our poll results differed greatly from the actual U.S. vote, the results were not significantly different ( $p$ -value=0.35).



Our results should be used with caution for several reasons. Our sample size was small, which can increase the sampling error, especially for estimates of subpopulations. Our exit poll results also may not be representative of our precinct. Even though our results were not statistically different from the actual precinct results, we did have a much higher proportion of voters who voted for Obama. There could be several explanations for this. Data were only collected during a two-hour period on Election Day, even though the polls were open for 13 hours. As different people tend to vote at different times of day, we likely missed people with different characteristics who voted in the morning or in the evening. For example, older people tend to start their days earlier and may have voted in larger numbers in the morning. People who work full-time also likely voted either before or after work.

The difference also could be due to nonresponse among mostly older white males, a group that generally supported McCain in the election. In the 2004 presidential election, exit polls overstated support for Kerry. One of the reasons given was that older Bush supporters had higher nonresponse rates. Even though only six older white males refused to participate in our exit poll, they could have made a difference in the results. Assuming all six refusals were McCain supporters (which is a big assumption), the proportion of McCain support in our poll would have risen from 23.3% to 37.9%, which is much closer the actual precinct vote outcome of 35.7%.

## Lessons Learned

All considered, we learned a lot about survey management by conducting this exit poll. Following are some of the most important points we learned:

It is essential to have a clearly defined objective about what the exit poll is trying to measure. Once the objective is clear, constructing the questionnaire becomes easier to do.

Pretesting the process is critical to maintain data quality on the day of the actual exit poll. Pretesting helped us determine how to divide the interceptor and recorder tasks and how to deal with unexpected situations. For example, we realized that not everyone leaving the poll is a voter. As a result, a “not eligible” category was added to our nonresponse sheet. Pretesting also allowed us to perfect our opening script and made us realize that conducting an exit poll was not such a scary task as it first appeared to be.

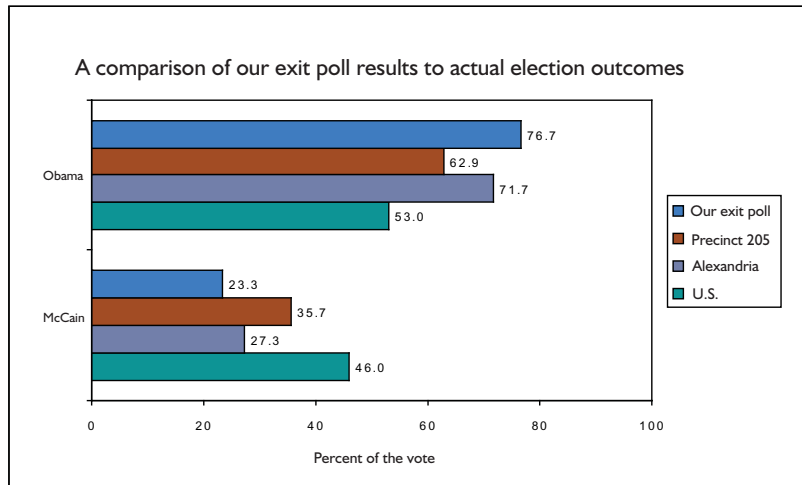


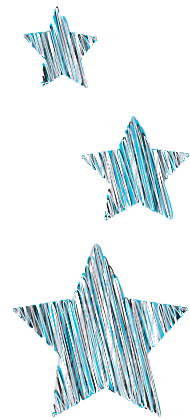
CHART 1. Our results show Obama won the presidential horse-race question with 76.7% of the vote.

A useful way to minimize nonresponse is by giving a clear introduction and explaining the purpose of the poll, by being friendly and talking clearly with respondents, and by providing a ballot box and making sure answers are confidential.

However, there are also points we did not consider that could ensure better data quality in the future:

More precise results can be obtained by increasing the sample size and collecting data throughout the day.

Bring a lot of clipboards. Having more than one clipboard can help reduce the number of missed respondents. At the beginning of our poll, we missed several respondents because voters exited in large numbers and we didn't have enough clipboards to hand out. Luckily, we had spare clipboards and were able to transfer questionnaires to the additional clipboards. Had we had additional clipboards in the beginning, our miss rate would have been much lower. ●



Editor's Note: Some of us will be talking about the 2008 election for a long time. This article should help us stay granular. Additionally, there may be elements students might want to model in a class project during the next election. For still more examples see [www.votingsystems.us](http://www.votingsystems.us) and the next issue of *STATS*

The views expressed are solely those of the authors and do not necessarily reflect the official positions or policies of the U.S. Government Accountability Office, where one of the authors now works.

# When Randomization Meets Reality

## *An impact evaluation in the Republic of Georgia*

by Miranda Berishvili, Celeste Tarricone, and Safaa Amer



The plastic bin was shaken seven times while the rattling of the plastic capsules inside it reverberated throughout the room. Hushed anticipation was palpable as a single capsule was drawn, the name inside it read aloud. The winner—an older Georgian farmer—rushed to the front of the room as the audience burst into applause. He spread his arms wide, and, in broken but determined English, shouted, “Thank you, America!”

The prize? A grant to expand his poultry operation. His effusive show of gratitude toward the United States was due to the grant being part of an agribusiness development project—known as ADA—funded by the Millennium Challenge Corporation (MCC), an American foreign aid agency that focuses on reducing poverty through economic growth.

The demand for ADA funds is far beyond the 250 grants (over four years) that the program will be able to issue. Recognizing a rigorous program evaluation opportunity as a result of this over-subscription, MCC decided to sponsor an experimental impact evaluation, conducted by NORC, to track the performance of program participants (the “treatment group”) against a statistically similar group of farmers (the “control group”). The treatment and control groups were selected through a randomization process.

Randomization, random allocation of the experimental units across treatment groups, is a core principle in the statistical theory of design of experiments. The benefit of this method is that it equalizes factors that have not been explicitly accounted for in the experimental design. While it is a simple, elegant solution for ensuring a statistically

rigorous experiment, the implementation of the process is complex, particularly when it comes to quality control.

Mechanical randomization, done by people rather than a computer, is even more challenging—whether flipping coins, selecting a card from a deck, or holding a lottery among Georgian farmers. Technical problems can arise easily, such as when the attempt at randomization broke down during the 1970 Vietnam draft lottery. In addition, resistance to the methodology among program implementers and potential program participants presents another obstacle. It is hard for these parties to accept that random selection will provide more rigorous results than subjective “expert judgment” in making a selection.

### **MCC: How and Why**

Established in 2004, MCC selects countries to receive its assistance based on their performance in governing justly, investing in their citizens, and encouraging economic freedom. One of the cornerstones of the corporation’s foreign aid approach is a strong focus on results. All programs have clear objectives and quantitative benchmarks to measure progress, and, whenever possible, projects are independently evaluated through rigorous methods to promote accountability among countries. In particular, program evaluations strive to establish causation between MCC projects and results by comparing the program against a counterfactual, or what would have happened to beneficiaries absent the program. These impact evaluations are used to garner lessons learned that can be applied to future programs and offered to the wider development community.

Georgia was part of the first group of countries selected for MCC funds. Its \$295 million program, which began in 2005, includes five projects covering agribusiness, private sector development, energy, and infrastructure. The ADA project focuses on small farmers and agribusiness entrepreneurs whose business plans have the potential to create jobs, increase household income, and foster growth in agribusiness—a promising sector of the Georgian economy.

Although MCC had identified the ADA project as a strong candidate for a rigorous impact evaluation,

program implementers in Georgia were not initially supportive of randomization. What won them over was that random selection lent credibility and transparency to the grant process—an important factor in a country with a long-standing history of corruption and skepticism of the state. Randomization offered a solution less vulnerable to selection problems and bias. The fundamental ethical rule in deciding whether to randomize is that when there is an oversupply of eligible recipients for scarce program resources, randomized assignment of candidates for the resource is fairer than relying solely on program scoring.

### Pilot Testing

Program implementers insisted, however, that the randomization be conducted publicly to promote transparency and understanding of the process. Due to the challenges of conducting such a public randomization event, pilot testing of the procedure and a few experiments were deemed essential. These experiments aimed to check the integrity of the procedure and ensure that the mechanics of stirring and shaking the containers achieved the random spreading of cases within the bin, thereby ensuring the quality of the selection process (i.e., that it was “random enough”). These pilot tests mirrored the process used during the actual randomization.

The following shows photos from two experiments to illustrate the pilot-testing process. These pictures show how the stirring and shaking took care of the original clustering that was intentionally arranged inside the bin. The process also allowed the implementers to check weak spots, such as corners, as observed in the photos from the first experiment.

## Experiment 1

### Step 1.

#### Clustered arrays

Yellow and brown containers were put into two groups within a plastic transparent bin. Note the brown clustered array was purposely placed in the lower right corner and two orange containers in the upper left corner.



1.

### Step 2. Stirring

The person on the left stirred the containers without looking in the bin, using a wooden spatula to mix the different colors, while the person on the right held the bin to stabilize it and prevent it from tilting.



2.



3.

### Step 3.

#### Result after stirring

The orange and brown containers were mixed, but the brown containers from the corner did not spread equally throughout the bin.

### Step 4. Shaking

The bin was covered with a tight transparent lid and shaken up and down, left and right, and then turned over.



4.

### Step 5.

#### Result after shaking

The brown containers were better spread throughout the bin as a result of shaking. Still, the implementers were not satisfied, so the process was continued in subsequent experiments, one of which is illustrated in Experiment 2.



5.

# Experiment 2



## Step 1. Larger clusters

This time, the two colors were clustered on each side of the bin, showing larger clusters compared to the previous experiment. The bin is fuller here, as we were testing to see if a larger bin might be needed.

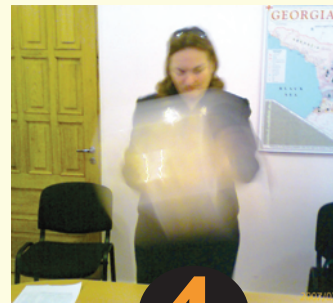
## Step 2. Stirring

Note that the bin tilted during the stirring. That reduced the effect of stirring and put some of the containers at risk of falling out of the bin. This was repaired in the actual selections.



## Step 3. Results after stirring

Note that the brown and yellow colors did not mix well as a result of stirring, perhaps because of the tilting.



## Step 4. Shaking

Different directions were used in shaking this time, including semicircular movements in and out.



## Step 5. Results after shaking

These results show a better mix of the colors, compared to the previous experiment.

*Note: In the actual randomization, all the containers were yellow.*

The repetition of the experiments with different scenarios and arrangements of colored containers in the bin allowed the implementers to improve the procedure, identify problems that could compromise the quality of the selection, and establish a method of mixing the containers that resulted in a wider spread inside the bin. The quality of the selection process is crucial, as it is a cornerstone of the quality of the impact evaluation.

## Implementation

During the first year of the program, three randomization events were held. The first two were relatively small, as the project was still ramping up. However, the third event, held in April of 2007, was considerably larger—14 grantees were selected among 36 applicants, with

more than \$300,000 in funds awarded.

The randomization was carefully moderated and narrated to address potential nervousness about the process. In addition, it made conscious use of key props to magnify the process: (a) equally sized pieces of paper with registration numbers, (b) homogeneous small plastic containers of the same color – one and only one for each eligible applicant, (c) a transparent rectangular bin, and (d) a wooden stick for stirring the containers in the bin.

To summarize, the application registration numbers were originally in a sequential list, ordered by the date of application receipt. This list was put into random order using a SAS random number generator. The moderator then read out the name of each applicant in this random order, and the applicant inserted a slip of paper with his registration number into one of the small plastic

containers, closed it, and put it into the transparent rectangular bin. This part of the process was carried in the same fashion until registration numbers for all the 36 eligible applicants were placed in the transparent bin. This enhanced the transparency of the process.

As each container was added, the containers in the bin were stirred seven times with a wooden stick by one of the facilitators. Then, the moderator put the lid on the bin and shook it at least seven times (up and down, left and right, and over the top).

One of the event participants—using another randomly sorted list—selected a container without looking into the bin. The chosen grantee was then identified to the audience. The process was repeated, each time stirring the containers, shaking the bin, and randomly selecting another individual from the audience to draw a name until all 14 grantees were chosen.

After all grantees were chosen, the same procedure was used to open the remaining containers to make sure each name appeared only once and to track the complete order of selection. Tracking the order of selection allowed running the statistical tests described in Table 2.

To ensure their quality, the random assignments were tested. In the case of a finite sequence of numbers, it is formally impossible to verify absolutely whether it is random, but it is possible to check that it shares the statistical properties of a random sequence, though even this can be a difficult task.

Note that numbers in a random sequence must not be correlated with each other or other sequences (i.e., knowing one of the numbers in a sequence must not help predict others in the same sequence or other sequences). Table 1 illustrates the correlation between three sequences: (1) the original sequence of receiving the applications, (2) the sequence of placing the containers in the bin (itself random), and (3) the sequence of selecting winners from the bin. This table shows that none of the bivariate correlations are significant.

To test for a serial correlation within the sequence of selecting winners, the Box-Ljung statistic was used. This test, like the previous one, was again not significant. The result of the Box-Ljung test is visually observed through the autocorrelation function (Figure 1).

One of the most common tests of randomness is the runs test. Results of the runs test for the selection sequence are shown in Table 2. These results confirm the initial findings from the autocorrelation function. Again, we fail to reject the hypothesis of randomness.

### Start on Inference

The scoring done by ADA program managers allowed us to split the original list of applicants into those eligible to receive a grant and applicants

who were not eligible to receive a grant. The randomization further split the eligible into a treatment group of grant recipients and a control group of eligible applicants who were not selected to receive a grant. Figure 2 shows box plots of the amount of money requested by applicants from different groups (right side), as well as the scores for 36 eligible applicants (left side).

Now with the treatments and controls identified, one of our next steps was to match businesses in each group pairwise using various factors such as program

		Original	Selection	Calling
Original	Correlation Coefficient	1	0.183	0.089
	Sig. (2 tailed)		0.285	0.604
	N	36	36	36
Selection	Correlation Coefficient	0.183	1	0.100
	Sig. (2 tailed)	0.285		0.561
	N	36	36	36
Calling	Correlation Coefficient	0.089	0.1	1
	Sig. (2 tailed)	0.604	0.561	
	N	36	36	36

TABLE 1. Spearman's Rho Correlation

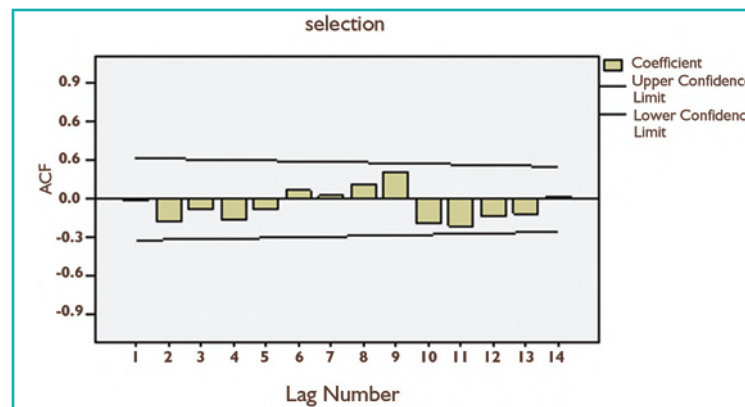


FIGURE 1. Autocorrelation function

	Based on Median	Based on Mean
Test Value	376	390
Cases < Test Value	18	20
Cases ≥ Test Value	18	16
Total Cases	36	36
Number of Runs	16	18
Z	-0.845	-0.095
Asymp. Sign. (2-tailed)	0.398	0.924

TABLE 2. Runs Test

## Poverty, War, and Statistics—Georgia This Time

by Safaa Amer and Fritz Scheuren

Are statisticians ever ready to deal with questions of war? Perhaps partly, but in the case of the Russian invasion of Georgia, our answer is “obviously not.” At least not yet. This is true, even for those of us who work/live in Georgia.

No matter how a war starts or is triggered, it always takes civilians by surprise. The war in Georgia was no different, even though we had been there just weeks earlier.

War’s impact has many dimensions. The first and foremost is the human factor. This human factor starts a chain reaction that affects not only individual lives, but goes beyond that to change economic trends and political attitudes.

Numbers are powerful in these situations. Our work on estimating the dead in Iraq due to the war showed the most striking fact is in the number of war-related casualties, as distinct from combat deaths.

Early media estimates of deaths occurring during the war in Georgia were around 4,000. At first, we were shocked by this number, but then we wondered who came up with this number and how? How much is simply an appeal for help? How much is factual? Now that some time has passed, we have a better estimate.

According to our colleague Mamuka Shatirishvili from Georgia, the actual number of deaths may be closer to 400. Still, that does not lessen the tragedy.

Estimating the number of dead and injured is the first consideration of damage by war. Buildings and infrastructure can be repaired, but the dead do not return to life. Their permanent loss is not only heart-breaking on a human level, but has strong economic and political effects, too. Statistics can be one way to assess these consequences and find ways to handle them.

We started our work in Georgia with an impact evaluation of poverty reduction through economic growth, thinking that was hard enough in a peacetime situation. Our client is the Millennium Challenge Corporation, an arm of the U.S. State Department. Problems with frames, survey instrument design, and data collection held more than a handful of challenges for us. But, now, the war with Russia has added a different dimension to program impact evaluation and widened our scope beyond the achievement of Millennium Development Goals. Just as the efforts for poverty reduction were starting to bloom in Georgia, everyone was surprised by the dramatic turn of events. Bottom line, what is our role as statisticians in this situation? Three partial answers come to mind, and we invite our colleagues to offer more.

To start with, we need to approach the problem from a human perspective, with sensitivity, and always maintain the rights of respondents when engaged in any data collection and analysis.

Next, we need to employ the most advanced methods our discipline has, seeking input, with all humility, from outside experts—especially those familiar with assessing the impact of other wars or natural disasters (such as Katrina).

Last, but not least, our goal is always to “keep our eyes on the prize,” as the Reverend Martin Luther King Jr. reminded us 45 years ago on August 28, 1963. His answer then was “freedom.” Ours should be that, too—the freedom to live a decent life in peace. Statistics must ever be to extend solidarity to our fellow humans who are suffering during this war and to offer them all respect and consideration.

*This article originally appeared in Amstat News, the membership magazine of the American Statistical Association, in October 2008.*

score and grant size, as exhibited above. There are small imbalances between the two groups that will have to be accounted for in the evaluation, knowing now that the imbalances are random.

### Next Steps

Refinements to the mechanical selection procedure will continue to be made, as well as improvements to the application and scoring processes, which will further reduce human bias and increase the precision

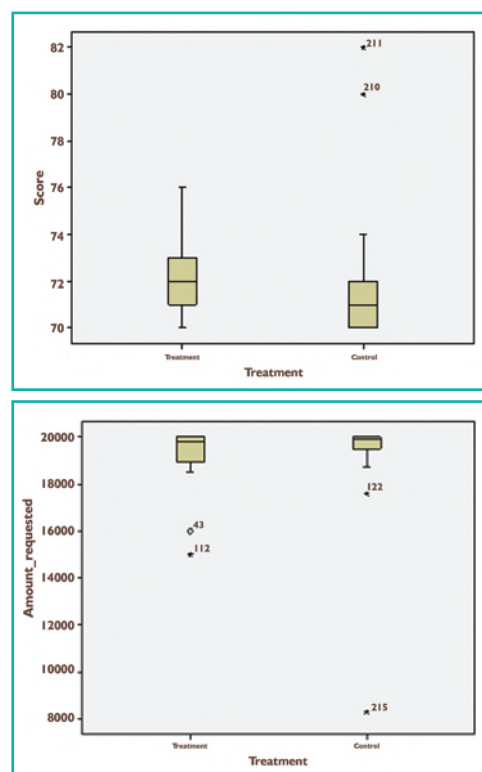


FIGURE 2. Box plots

of the impact evaluation. With the randomization process relatively stable, attention will now turn to data collection of the treatment and control groups and analysis of results. As the first year of the ADA project came to a close, the use of randomization met with considerable success and, finally, acceptance. This method of selection will continue to be used during the remaining three years of the program. As one of the Georgian program implementers put it, “This process has never been used in Georgia before, for any kind of program.” Now that it has attracted interest and support, perhaps it will be used more often.

The views expressed here are those of the authors, and do not necessarily represent the views of the Millennium Challenge Corporation. The content contained herein was approved for release by the Millennium Challenge

Editor’s Note: *Randomization is well-developed, but its application, in a manner that requires full transparency, can be a challenge. See Poverty, War, and Statistics—Georgia This Time.*

# Matching Items and Collecting Coupons



by Bruce E. Trumbo and Luther B. Scott

In this column, we explore two quite different kinds of matching problems. Some of the results can be obtained analytically, but mostly we use simulation. You may recall the article about birthday matches in *STATS* Issue 43, Spring 2005. If so, you will recognize that some of the simulation methods used here are similar to those in that article.

## The Item-Matching Problem

A few years ago in her weekly column for *Parade* magazine, Marilyn vos Savant presented the following scenario: “A high-school student who hadn’t opened his American history book in weeks was dismayed to walk into class and be greeted with a pop quiz. It was in the form of two lists, one naming the 24 presidents in office during the 19th century in alphabetical order and another list noting their terms in office, but scrambled. The object was to match the presidents with their terms. The completely clueless student had to guess every time.” Then she went on to ask for the expected number of questions the student gets right.

Intuitively, the answer is one, because the student has one chance in 24 of guessing the correct answer for any of the 24 presidents. More formally, let  $X_i$  equal 0 if the student gives the wrong answer on the  $i^{\text{th}}$  question and 1 for the correct answer. Then  $P\{X_i = 1\} = E(X_i) = 1/24$ . The total number of correct answers or matches is  $S = X_1 + X_2 + \dots + X_{24}$ , and so

$$\begin{aligned} E(S) &= E(X_1 + X_2 + \dots + X_{24}) \\ &= E(X_1) + E(X_2) + \dots + E(X_{24}) = 24(1/24) = 1. \end{aligned}$$

Here, there are  $n = 24$  items to match, but if you look at this computation carefully, you will see that  $E(S) = 1$ , regardless of the number of items there are to match. Some people are surprised that the expected number correct does not increase as  $n$  increases.

## Simulating the Item-Matching Problem

It is not difficult to simulate this situation, and we can learn some interesting facts about the scores on item-matching tests by doing so. The heart of the simulation lies in taking an ordered random sample of size 24 without replacement from a population of 24 items. This gives a random permutation of the 24 items, which amounts to a ‘randomly guessed’ set of answers to the test. The sampling can be simulated in R with the code `perm = sample(1:24, 24)`. One run of this code gave the following result:

```
> perm = sample(1:24, 24)
> perm
[1] 17 15 22 16 23 3 20 18 24 9 11 14
[13] 2 13 10 1 5 4 12 6 8 19 21 7
```

Next, we need to find out how many of the 24 numbered items remain in their correct location in sequence after permutation, which is the exam score  $S$ . If you count across the result above, you will see that item 11 is in the eleventh position and that it is the only such match. So, the simulated value of  $S$  is one.

The R code `s = sum(perm==1:24)` performs this count automatically. The logical vector resulting from the code `perm==1:24` consists of 24 elements: In our example, its eleventh element is TRUE because the eleventh element of `perm` is 11. All other elements of this logical vector are FALSE. The sum of this vector treats TRUE as 1 and FALSE as 0. Thus, the value of the number of matches `s` below is 1.

```
> s = sum(perm==1:24)
> s
[1] 1
```



**LUTHER SCOTT** completed his MS in statistics at California State University, East Bay in 2007. He is currently a senior analyst at Kaiser Permanente in Oakland, California. In his spare time, he enjoys hiking and running.



**BRUCE TRUMBO** is a professor of statistics and mathematics at California State University, East Bay (formerly CSU Hayward). He is a Fellow of the American Statistical Association and a holder of the ASA Founder’s Award.

By looping through these two statements  $m = 100,000$  times and making a vector  $s$  of the  $m$  scores that result, we can simulate the distribution of the random variable  $S$ . The full program is shown in Figure 1; the numerical results of one run are shown below the program, and the histogram of the simulated distribution of  $S$  appears in Figure 2.

```
m = 100000
# Number of iterations
n = 24
# Number of items to match
s = numeric(m)
for (i in 1:m) {
  perm = sample(1:n, n)
  # Random permutation
  s[i] = sum(perm==1:n) }
mean(s); sd(s)
cutp = 0:(n+1)-.5
# Needed for a nice histogram
hist(s, breaks=cutp, prob=T)

> mean(s); sd(s)
[1] 0.99918
[1] 1.001623
```

FIGURE 1. R code and numerical results for 100,000 randomly answered item-matching tests with 24 items

In this program, `mean(s)` approximates  $E(S)$  and `sd(s)` approximates  $SD(S)$ . Notice that both values are very nearly 1. Actually,  $V(S) = 1$ , regardless of the value of  $n$ . The proof for the variance is not quite as easy as the proof for the expectation in the previous section. It is more difficult because our  $X_i$  are not *independent*. (When random variables are independent, the variance of their sum is equal to the sum of their variances, but otherwise not necessarily equal.) For example, it is not possible for all but one of them to be equal to 1, which is the same as saying  $P\{S = n - 1\} = 0$ .

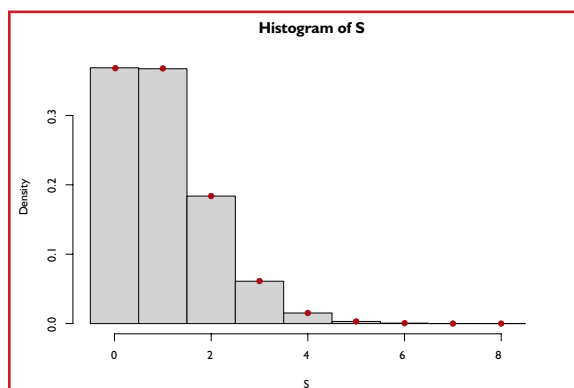


FIGURE 2. Histogram of scores  $S$  on an item-matching test with 24 items. Possible values extend out to 24, but probabilities of scores larger than about 5 are too small to show on the graph.  $E(S) = SD(S) = 1$ , no matter how many items are to be matched. The heavy dots show probabilities from  $POIS(\lambda = 1)$ .

If a random variable is distributed according to a Poisson distribution with mean  $\lambda$ , then its variance is also equal to  $\lambda$ . So, we wonder if there may be some relationship between the distribution of scores  $S$  on a randomly answered item-matching test and the Poisson distribution with  $\lambda = 1$ . The answer is that  $S$  is very nearly distributed  $POIS(1)$ , especially for moderately large or large values of  $n$  (say  $n \geq 10$ ). The dark dots in Figure 2 show exact probabilities for  $POIS(1)$ .

However, the distribution of  $S$  is not *exactly* Poisson. We saw above that  $P\{S = n - 1\} = 0$ . Also,  $P\{S = n\} = 1/n!$ , and  $P\{S = k\} = 0$ , for all  $k > n$ . None of these statements is precisely true for a random variable  $X$  distributed as  $POIS(1)$ , for which  $P\{X = k\} = 1/ek! > 0$ , for all  $k = 0, 1, 2, \dots$ .

Only one of our results about the distribution of the score  $S$  on an item-matching test was easy to obtain by analytic means [that  $E(S) = 1$ ], and we have been able to verify this result with simulation. But more important, with simulation, we also have been able to guess that  $V(S) = 1$  and to show that the distribution of  $S$  is well-approximated by  $POIS(1)$ . These results are not so easily proven analytically.

This item-matching problem is a famous one in probability theory. It has appeared in several disguises.

One is that  $n$  men check their hats at a banquet, each receiving a numbered receipt to claim his hat when the banquet is finished. But the person in charge of the hats accidentally scrambles the corresponding numbers on the hats and passes them back at random. What is the expected number of men who get their own hats back? (This is from the days when gentlemen customarily wore hats, but only out of doors.)

Another is that a very tired administrative assistant has  $n$  letters and  $n$  envelopes with matching names and addresses, but scrambles them at random when it comes time to put letters into envelopes. What is the expected number of recipients who get the letter intended for them?

A general way to phrase the problem in mathematical terms is to ask for the expected number of points that remain fixed under a random permutation of points.

## The Coupon-Collecting Problem

This matching problem is based on a particular kind of promotion intended to increase sales of various products, often boxes of food or bottles of beverage. The manufacturer places one of  $n$  types





of coupons at random in each box or under the cap of each bottle. Suppose the types are numbered from 1 through  $n$ . When the customer is able to collect one of each of the  $n$  types of coupons, he or she wins a prize.

Let's begin by assuming all  $n$  types of coupon are equally likely, although that is not usually the case in practice. Here, the main random variable of interest is the number  $W$  of coupons a customer needs to collect to get a complete set of all  $n$  types of coupon.

An argument based on the geometric distribution allows us to find  $E(W)$  and  $V(W)$ . Suppose we perform independent binary trials (Bernoulli trials) with possible results  $S$  and  $F$ , having  $P(S) = p$  and  $P(F) = 1 - p = q$  on each trial. Let  $X$  be the number of the trial on which  $S$  appears for the first time. One can show that  $E(X) = 1/p$  and that  $V(X) = q/p^2$ .

Of course, the first coupon collected cannot be a duplicate, so the waiting time  $X_1$  for the first useful coupon is deterministic:  $X_1 = 1$ . The additional waiting time  $X_2$  for the second useful coupon is geometric with  $p = (n - 1)/n$ , because any coupon that does not match the first is useful. Similarly, the additional waiting time  $X_3$  (number of purchases beyond  $X_2$ ) for the third useful coupon is geometric with  $p = (n - 2)/n$ . And finally, the additional waiting time  $X_n$  for the last and winning coupon is geometric with  $p = 1/n$ .

In this notation,  $W = X_1 + X_2 + \dots + X_n$ . Thus, using an argument similar to that above—that the expectation of a sum is the sum of the expectations—we have

$$E(W) = E(X_1) + E(X_2) + \dots + E(X_n) = 1 + n/(n - 1) + \dots + n.$$

For any particular value of  $n$ , it is easy to evaluate this in R. For example, if  $n = 6$ , then `sum(n / (1 : n))` returns  $E(W) = 14.7$ . Here, we can evaluate  $V(W)$  as the sum of variances  $V(X_i)$  because the  $X_i$  are based on non-overlapping sets of independent purchases and are therefore independent. (See Challenge 6 at the end of this article.) Figure 3

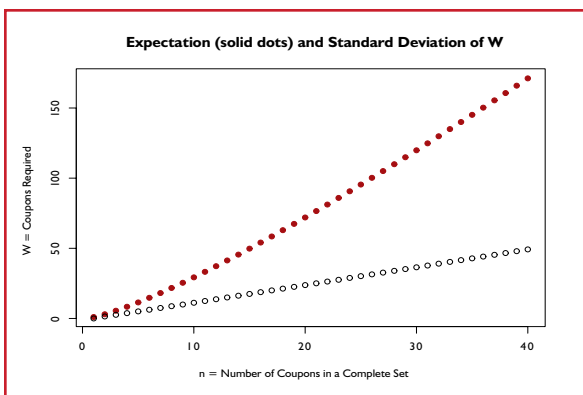


FIGURE 3. In the coupon-collecting problem, both  $E(W)$  and  $SD(W)$  increase with increasing  $n$ .

shows values of  $E(W)$  and  $SD(W)$  for  $n = 1, \dots, 40$ .

It is fairly easy to get a couple of particular values in the distribution of  $W$  by analytic means. For example, suppose  $n = 6$ :

The smallest possible value of  $W$  is six. Here, the customer is lucky enough to get a complete set of six types of coupons on the first six purchases:

$P\{W = 6\} = 6!/6^6 = 0.0154321$ . (This is the same as the probability of seeing all six faces in six rolls of a fair die.)

Finding  $P\{W = 7\}$  is a little more difficult.

Obviously, the denominator is  $6^7$ . In the numerator, we have to account for one redundant coupon: six possibilities. Then, we have to choose on which two of the first six purchases the two matching coupons were purchased: 15 possibilities. Finally, we have to account for the arrangements of the five unique coupons:  $5! = 120$  possibilities. So

$$P\{W = 7\} = 6(15)(5!) / 6^7 = 0.03858025.$$

In this case with  $n = 6$ , simulation gives the following approximate values:  $E(W) = 14.77$ ,  $P\{W = 6\} = 0.015$ , and  $P\{W = 7\} = 0.039$ . The method of simulation is explained below; these values are based on 10,000 simulated collectors. Figure 4 shows the simulated distribution of  $W$  when  $n = 6$ . The histogram is plotted on a probability scale so it shows, within the accuracy of reading the graph, the approximate values of  $P\{W = 6\}$  and  $P\{W = 7\}$  along with the other values in the distribution of  $W$ .

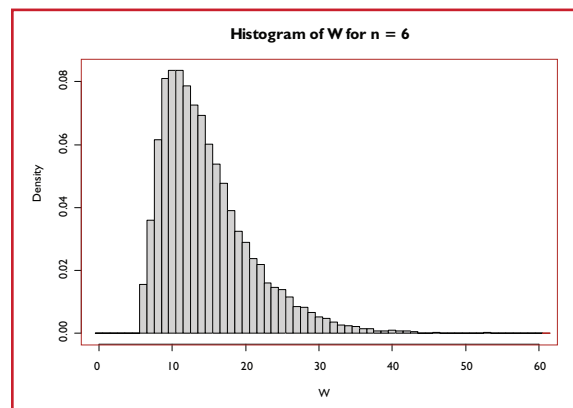


FIGURE 4. Simulated distribution of the number  $W$  of coupons one needs to collect to get a complete set of six equally likely coupons. Alternatively,  $W$  can be viewed as the number of rolls of a fair die required before all six faces are seen.

Those who test random number generators often simulate the coupon-collecting problem as a benchmark. When some kinds of flawed random number generators are used to simulate the

distribution of  $W$ , they give outrageously wrong answers. (See “R U Simulating? ‘Random’ Numbers from Nonrandom Arithmetic” in *STATS* Issue 46, Fall 2006.)

### Simulating the Coupon-Collecting Problem

Now we explain the method we used to simulate the distribution of  $W$ . To simulate a particular realization of  $W$ , we use a `while`-loop. As we begin, the vector `got` is an initial collection of  $n$  coupons. That may be—but probably will not be—enough for a complete set. If it is not enough, we “buy” one more coupon on each passage through the loop until we have a complete set; that is, when `u` finally equals the number of coupon types in a complete set.

```
got = sample(1:n, n, repl=T)
u = length(unique(got))
while (u < n){
  got = c(got, sample(1:n, 1))
  #get another
  u = length(unique(got)) }
w = length(got)
```

When the function `unique` is applied to a vector, the redundant elements of the vector are removed. For example, after one run of the code above with  $n = 6$ , we obtained the following results based on the final values of `got` and `u`. In this instance, it was necessary to buy nine coupons in addition to the original six to get a complete set of six types.

```
> got; unique(got); u; w
[1] 6 3 2 1 5 5 3 5 5 6 6 3 5 5 4
[1] 6 3 2 1 5 4
[1] 6
[1] 15
```

In the case where there are  $n = 20$  coupons to collect, we used an outer `for`-loop to iterate the above procedure 10,000 times and, thus, to make a vector `w` of 10,000 waiting times  $W$ . Here are our results from one run, obtained with the additional code `mean(w); sd(w); mean(w > 100)`:

- $E(W) \approx 71.9$  (exact value 71.955)

- $SD(W) \approx 24.1$  (exact value 23.801)
- $P\{W > 100\} \approx 0.115$  (exact value unknown)

Because the simulated mean and standard deviation are close to the known exact values, we have some confidence that we can also rely on the simulated value  $P\{W > 100\}$  to be reasonably accurate. Figure 5 shows the histogram of the simulated distribution of  $W$ .

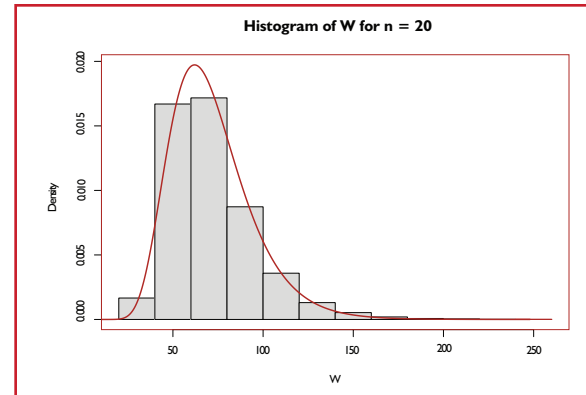


FIGURE 5. Simulated distribution of  $W$  for  $n = 20$ , with an approximating lognormal density curve

The lognormal distribution is a right-skewed distribution. If  $X$  is lognormal, then  $Y = \ln(X)$  is normal, say  $NORM(\mu, \sigma)$  to be specific. (We are using natural logarithms.) Then, it is customary to use  $\mu$  and  $\sigma$  as the parameters of the lognormal distribution, even though they are not the mean and standard deviation of  $X$ . It turns out that for moderately large  $n$ , the number  $W$  of coupons needed to get a full set is approximately lognormal. The lognormal density curve in Figure 5 was drawn using the simulated values  $\mu = E[\ln(W)] \approx 4.22$  and  $\sigma = SD[\ln(W)] \approx 0.311$  as parameters.

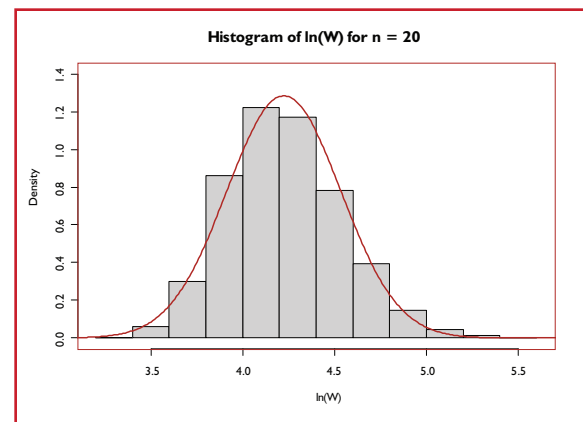


FIGURE 6. Logarithms of the values in Figure 5 are plotted.



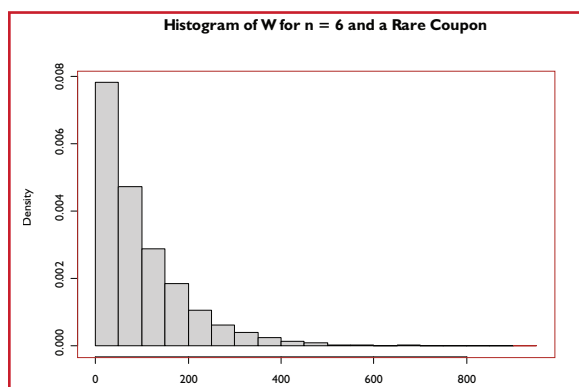


FIGURE 7. Distribution of the number of coupons required to get a full set when one of the coupons appears only one time in 100 purchases

The lognormal approximation seems to be reasonably good, except for very small  $n$ . For  $n = 20$ , the sample quantiles 0.10, 0.25, 0.50, 0.75, 0.90, and 0.95 of the simulated distribution of  $W$  are 47, 55, 67, 84, 103, and 116, respectively. The logarithms of these sample quantiles fall at the quantiles 0.11, 0.24, 0.48, 0.75, 0.91, and 0.96 of the distribution  $\text{NORM}(4.22, 0.311)$ . This means, for example, that if  $Y \sim \text{NORM}(4.22, 0.311)$ , then  $P\{Y \leq \ln(84)\} = 0.75$ . Figure 6 shows log-transformed values of Figure 5.

Our simple lognormal approximation is based on matching the mean and variance of  $W$ . We mentioned at the start of our discussion of the coupon-collecting problem that, in an actual advertising promotion, the types of coupons are usually not equally likely. In the case where  $n = 6$ , suppose that the probabilities of coupon types one through five are each 0.198 and that the probability of coupon six is only 0.010. In this extreme circumstance, we are very likely just waiting to get a coupon numbered six to complete the set, and so the distribution of  $W$  must be nearly geometric. This ‘unfair’ situation will probably have to be explained in the very fine print that announces the promotion. And there will probably be some ads on eBay for coupon six.

The `sample` function allows for sampling with these unequal probabilities by using the additional argument `prob=c(rep(.198, 5), .01)`. Figure 7 shows the result of simulating  $W$  for these unequal probabilities. This simulation also gives the approximate values  $E(W) = 102.0$  and  $SD(W) = 100.6$ . The mean and standard deviation of a geometric random variable with success probability  $p = 0.01$  are 100 and  $(0.99)^{1/2} / 0.01 = 99.5$ , respectively. ●

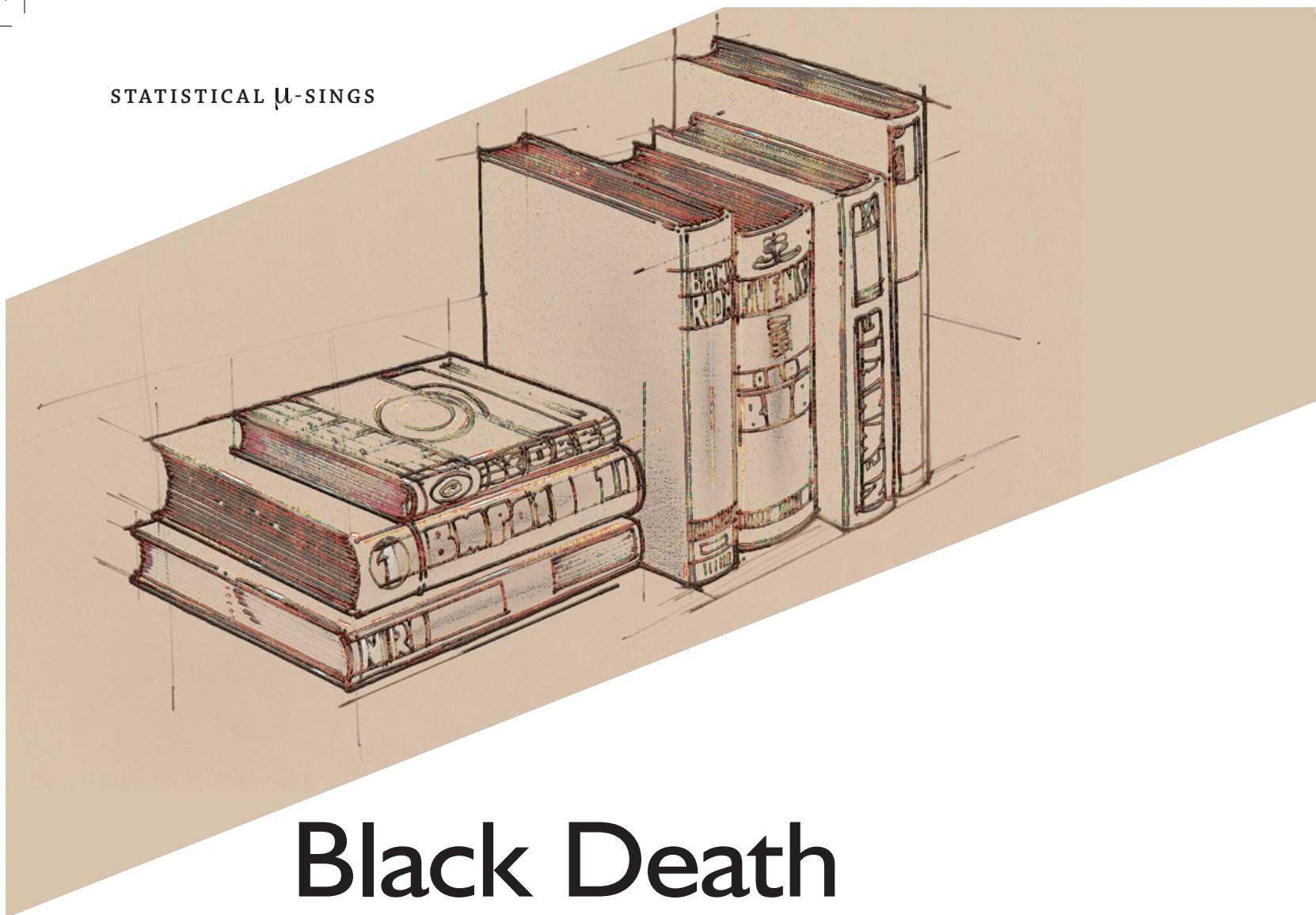
Editor’s Note: *This article is the most mathematical of this issue. The subtle blend of simulation, computing, and sound graphics has a fully modern feel.*

## CHALLENGES

We hope you will follow through with what we have done and then try to answer the following questions.

1. For an item-matching test with  $n = 4$  items, explicitly list the 24 possibilities and give the exact distribution of the number  $S$  correct. Verify directly from the distribution that  $E(S) = V(S) = 1$ .
2. Suppose you tried the method in Challenge 1 for  $n = 20$ . If you were able to write one item in the list each second, how long would it take you to make the list?
3. For an item-matching test with  $n$  items, give a logical explanation why  $P\{S = n - 1\} = 0$ .
4. For an item-matching test with 20 items, use the Poisson approximation to approximate the probability of getting fewer than three right answers. Simulate this probability and compare results.
5. You roll a fair die seven times. Use a combinatorial method to find the probability that you will see all six faces of the die. Why is this *not* the same as  $P\{S = 7\}$  computed earlier in the article? How is this related to the distribution of  $S$ ?
6. In collecting a set of 10 coupons, use the argument suggested earlier in this article to show that the standard deviation of the number of coupons you need to buy is 11.2110. Can you write R code to make Figure 3?
7. Suppose the six coupons in a coupon-collecting situation have probabilities  $1/21$ ,  $2/21$ ,  $3/21$ ,  $4/21$ ,  $5/21$ , and  $6/21$ , respectively. Make a figure similar to Figure 7 and give an approximate value for  $E(W)$ .
8. (Advanced) In the item-matching problem, show that  $V(S) = 1$ , regardless of the value of  $n$ . Hint: Begin by arguing that  $E(X_i X_j) = (1/n)[1/(n - 1)]$ , for  $i \neq j$ .

To check your answers, visit the *STATS* web site at [www.amstat.org/publications/stats](http://www.amstat.org/publications/stats).



# Black Death

## A review of *Applied Spatial Statistics for Public Health Data*



**CHRIS OLSEN** teaches mathematics and statistics at Thomas Jefferson High School in Cedar Rapids, Iowa. He has been teaching statistics in high school for 25 years and has taught AP statistics since its inception.

by Chris Olsen

As it happens, I was born and raised in the great state of Iowa, home of the quadrennial Iowa caucuses. As can only be appreciated by, perhaps, the citizens of the great state of New Hampshire, the months surrounding the 2008 presidential election were a time of plague—reaching near biblical proportions. We were a little light on the locusts and rivers turning to blood, but in the rate of phone calls, commercials, entreaties for support, and political mailings, we were sitting about five standard deviations above the mean.

It comes as no surprise, then, that my thoughts might have turned to more pleasant considerations, as a sort of defense mechanism. True, the elevation of my spirit could have been just some sort of placebo effect, but in times of

travail, I tend to find myself grasping at even the smallest hope. In this case, my ‘Shawshank Redemption’ kicked in when I recalled an introductory sentence in Howard Wainer’s book *Graphic Discovery*. “Let me,” he wrote, “begin with a few kind words about the bubonic plague.”

I hasten to add that I didn’t really remember this while watching political commentary. No, the phrase returned to me while I was thumbing through recent issues of the *American Journal of Physical Anthropology* in search of some chi-square problems with which to plague—as it were—my students on future tests. Lo and behold, I stumbled upon an article, “Paleodemographic Comparison of a Catastrophic and an Attritional Death Assemblage.” This wasn’t about

19th-century Iowa caucuses, but a comparison of two cemeteries in England: a ‘normal’ cemetery named St. Helen-on-the-Walls and a so-called “Black Death” cemetery, a little northeast of the Tower of London, over which the Royal Mint was built.

The statistical analysis part caught my eye, of course, and before I knew it, I was immersed in the world of paleodemography—whatever that is—in search of more chi-square problems. Just as I began to fathom the differences among bubonic, septicemic, and pneumonic plague, one of the references led me to a book by Lance Waller and Carol Gotway titled *Applied Spatial Statistics for Public Health Data*. One of the running data sets in this book was the analysis of medieval grave sites, and, of course, I started looking for more chi-square problems. (By that time, my tunnel vision had convinced me cemeteries were veritable beehives of categorical data.)

While leafing through the book, I realized I had stumbled across an area of statistics I really had not paid much attention to: spatial statistics. I had lots of leaf time to realize this because I didn’t find any mention of Pearson’s chi square until Page 242, and by that time, I had noticed lots of interesting uses of spatial statistics, some almost up to the interest level of medieval cemeteries. Of course, applying spatial statistics turned out to be too late for the folks in St. Helens-on-the-wall and northeast of the Tower of London, but it turns out that it isn’t too late for the study and interception of more modern maladies, with the possible exception of plagues of political types.

Fortunately for me, Waller and Gotway assume a minimal knowledge of spatial statistics in their presentation. Unfortunately for me, the assumption of no knowledge would have been better. Be that as it may, the authors present a plethora of statistical methods for addressing some very interesting questions. How would one define the concept of randomness applied to points in space, and how would nonrandom ‘clusters’ of points be detected? Once nonrandom clusters are detected, when do they become evidence of an outbreak of disease? How are such clusters associated with environmental hazards such as toxic waste dumps? These questions are, of course, important in epidemiological work, where the detection of outbreaks of disease and other health risks can save and extend lives.

As I have more than a passing interest in things epidemiological, I very much appreciated that the introductory chapters about analyzing public health data provided a bit of review and a bridge into the world of spatial statistics. Some of the methods

“

One of the running data sets in this book was the analysis of medieval grave sites, and, of course, I started looking for more chi-square problems.”



discussed in the book could have used a press agent—“K functions,” “L plots”—however, other, more interesting, names for statistical techniques, such as “headbanger smoothing,” made up for the singularly alphabetical methods.

Being a more or less casual reader of things statistical, as distinguished from being an impending epidemiologist, what I really appreciated about this book is that it expanded my horizons about the reach of statistical analysis with clear expositions of the problems addressed with spatial statistics. I also liked the especially interesting data sets used throughout the book. Of course, the authors could have put in more chi-square problems, but on the whole, they made up for this deficiency with a lively expository style. All in all, I would recommend picking up this book if you are interested in serious epidemiology or the discipline of statistics in general. ●

---

Editor’s Note: *Even though the 2008 elections are over, this ‘look back’ article might give you the feel of that “happening.” Anyway, it does not hurt to have the treatment be tongue-in-cheek.*

## 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 numbers in blue are the page numbers for each article.

### 3 *FiveThirtyEight.com*

More about Nate Silver and 538 at [www.fivethirtyeight.com](http://www.fivethirtyeight.com)

### 10 Designing and Implementing an Election Day Exit Poll

To read the official 2008 Alexandria precinct voting results, visit [www.voterinfo.sbe.virginia.gov/election/DATA/2008/07261AFC-9ED3-410F-B07D-84D014AB2C6B/Unofficial/00\\_p\\_510\\_89BE12EC-7BBF-479C-935A-9B8C51DD3524.shtml](http://www.voterinfo.sbe.virginia.gov/election/DATA/2008/07261AFC-9ED3-410F-B07D-84D014AB2C6B/Unofficial/00_p_510_89BE12EC-7BBF-479C-935A-9B8C51DD3524.shtml).

Edison Media Research and Mitofsky International (2008). Excerpt from *Evaluation of Edison/Mitofsky Election System 2004, Election and Exit Polling*, Scheuren, F. and Alvey, W. (eds.), Hoboken, NJ: John Wiley & Sons, 66–68.

Merkle, D.M., and Edelman, M. (2002). *Nonresponse in Exit Polls: A Comprehensive Analysis*, Survey Nonresponse, Groves, R.M.; Dillman, D.A.; Eltinge, J.L.; and Little, R.J.A. (eds.), New York, NY: John Wiley & Sons, 243–257.

Scheuren, F. and Alvey, W. (2008). *Introduction, Election and Exit Polling*, Scheuren, F. and Alvey, W. (eds.), Hoboken, NJ: John Wiley & Sons, 1–15.

### 16 When Randomization Meets Reality An impact evaluation in the Republic of Georgia

Bayer, D. and Diaconis, P. (1992). “Trailing the Dovetail Shuffle to Its Lair.” *Annals of Applied Probability*, 2(2):294–313.

Boruch, R. (1997). *Randomized Experiments for Planning and Evaluation: A Practical Guide*, Applied Social Research Methods Series, Thousand Oaks, CA: Sage Publications.

Duflo, E. and Kremer, M. (2003). “Use of Randomization in the Evaluation of Development Effectiveness.” Prepared for the World Bank Operations Evaluation Department (OED), Conference on Evaluation and Development Effectiveness in Washington, DC. <http://econ-www.mit.edu/files/765>.

Rosenbaum, P. (2002). *Observational Studies*. Springer Series in Statistics, New York, NY: Springer.

Starr, N. (1997). “Nonrandom Risk: The 1970 Draft Lottery.” *Journal of Statistics Education*, 5(2). [www.amstat.org/publications/JSE/v5n2/datasets.starr.html](http://www.amstat.org/publications/JSE/v5n2/datasets.starr.html).

The American Association for Public Opinion Research (2008). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*, 5th edition, Lenexa, Kansas: AAPOR.

### 21 Matching Items and Collecting Coupons

Birthday Matches, “R U Simulating?” *STATS* Issue 43. Spring 2005.

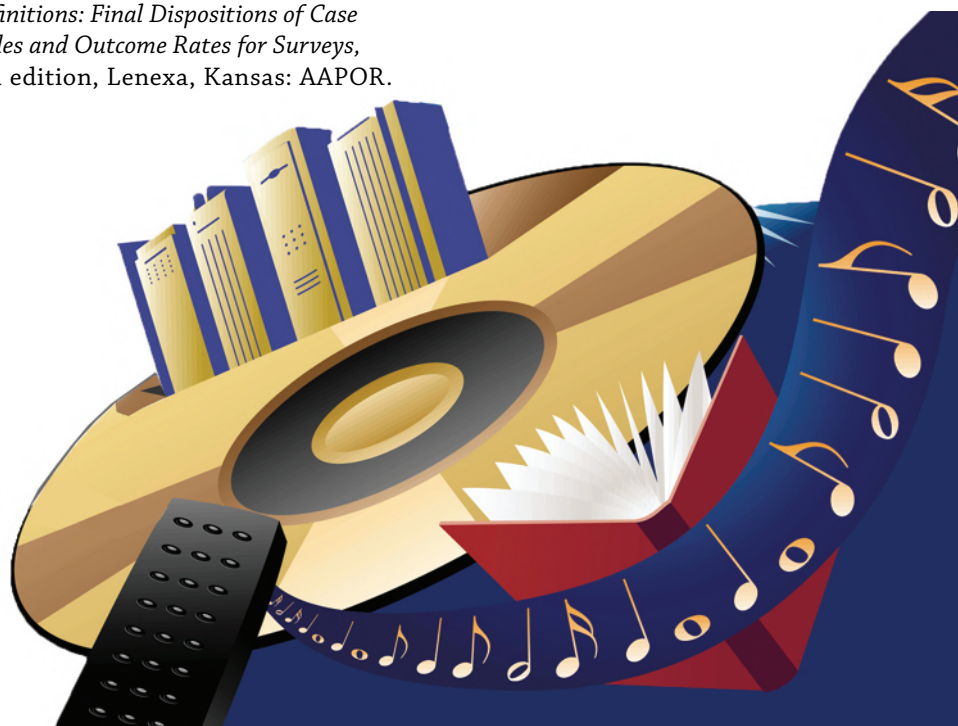
‘Random’ Numbers from Nonrandom Arithmetic, “R U Simulating?” *STATS* Issue 46, Fall 2006.

### 26 Black Death A review of *Applied Statistics for Public Health Data*

Margerison, B.J. and Knusel, C.J. (2002). “Paleodemographic Comparison of a Catastrophic and an Attritional Death Assemblage.” *American Journal of Physical Anthropology*, 119(2):134–143.

Wainer, H. (2005). *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*. Princeton, NJ: Princeton University Press.

Waller, L.A. and Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*. Hoboken, NJ: John Wiley & Sons.



# Join *now*



Become a Student Member  
of the **ASA** for only **\$10**

**Join the more than 4,000 students who  
already know what ASA membership means...**

**Free subscriptions** to *Amstat News*, your monthly membership magazine!

**Free online access** to the *Journal of the American Statistical Association*, *The American Statistician*, and the *Journal of Business & Economic Statistics*.

**ASA members-only discounts** on ASA publications, meetings, and Continuing Education Courses, PLUS special discounts from publishers.  
**A network of professional colleagues**

made up of more than 18,000 ASA members, including 4,000 students.

**Free or discounted dues** for most regional chapters and special-interest sections.

**Career opportunities and information** through [www.amstat.org](http://www.amstat.org), our JSM Career Placement Service, and *Amstat News*.

*[www.amstat.org/join](http://www.amstat.org/join)*



# The ASA Brings the Library to Your Desktop



Did you know your ASA membership includes online access to the:  
*Journal of the American Statistical Association*  
*Journal of Business & Economic Statistics*  
*Statistics in Biopharmaceutical Research* and  
*The American Statistician*?

Log in to MEMBERS ONLY today to access your journals!  
[www.amstat.org/membersonly](http://www.amstat.org/membersonly)