Sample selection and estimation in the Daybreak Poll

Erik Meijer

September 18, 2016

1 Recruitment for the poll

The Daybreak Poll is administered in the Understanding America Study (UAS). The UAS is a panel that is recruited through address-based sampling. Through our vendor MSG, we obtain addresses from the USPS Delivery Sequence File, and names and other contact details if MSG has them in their database. We then send an initial paper survey to these addresses. Respondents can indicate whether they are willing to participate in further studies. Those who agree are given login instructions to enroll in the panel. If they don't have internet access, the UAS provides them with access (tablet and/or internet connection), so we do not exclude individuals without internet access. The UAS webpages discuss sampling and recruitment in more detail, see `https://uasdata.usc.edu/recruitmentoverview/1`.

Individuals have to be at least 18 years old to be eligible for the panel. All household members of the initial respondent who are at least 18 years old are eligible to enroll in the panel, although the number of additional household members who enroll in the panel has so far been limited.

The UAS is mainly intended to be a nationally representative panel, but there are two subpanels with more narrow scope: a subpanel of Native Americans and a subpanel for families with young children in Los Angeles County. Note that the nationally representative panel also contains Native Americans and LA County residents. The additional subpanels are intended for specific analyses of these populations.

The initial recruitment for the UAS started in 2014, and we have been recruiting and adding panel members continuously since then. The panel is still growing. As of early September 2016, the UAS has about 5,500 panel members.

In May 2016, all panel members who are U.S. citizens were asked to respond to the pre-election survey UAS47. This asks a few questions about voting in 2012 and 2014, the subjective likelihood of voting in the 2016 election, and asks respondents whether they would be willing to participate in the poll. New UAS panel members are asked to complete UAS47 as one of their first surveys. Those who agree to participate in the poll constitute the *poll panel*. They are invited once a week to answer the three poll questions and the two additional questions that differ from week to week.

2 Assignment of the day of the week

Each member of the poll panel is invited for the poll once a week. We split the poll panel into seven groups and assign each group to a day of the week. Members receive the invitation to participate each week on the same day of the week, but they are allowed to respond up to six days later (the day before the next invite).

The first group assignment, based on all responses to UAS47 up to that time, was done on Saturday July 2, 2016. The second group assignment, using new responses to UAS47, was done on Saturday, July 16, 2016. Since then, each Saturday, new respondents to UAS47 who are willing to participate are assigned to a weekday.

The assignments are done in a way that intends to balance the characteristics of the members across the days, so that the samples for the different weekdays are similar. Because we aggregate results across a whole week (see section 3 below), this is not essential, but it improves stability and guards against anomalous results that might occur if some event shifts opinions of different subgroups differently. We first split the newly assigned poll members into core and non-core members. Core members are members from the nationally representative part of the panel that have no missings on any of the variables that are used to construct the weights in the poll (gender, race/ethnicity, education, age, family income, household size, voting in 2012). Non-core members are members who are from the Native American or LA County subpanels or who have missings on at least one of the aforementioned variables. Core members will get a positive weight in the poll and thus will be used to compute our results, whereas non-core members are not included in the poll sample (see section 3 below). We balance both groups separately in each weekly assignment.

We consider the following variables for the balanced assignment, loosely based on the variables that are used in the weighting procedure: (1) gender (0=male, 1=female), (2) race/ethnicity (1=non-Hispanic white alone, 2=other), (3) education (1=less than high school, 2=high school graduate or some college, 3=college graduate and above), (4) age (1=18–34, 2=35–54, 3=55+), (5) family income (1=less than $35,000, 2=$35,000–$74,999, 3=$75,000 or more), (6) household size (1=1, 2=2–3, 3=4 or more), (7) voting in 2012 (1=too young, 2=old enough but did not vote, 3=voted for Romney, 4=voted for Obama, 5=voted for someone else).

For the core sample, 7-dimensional cells are created from a complete cross-classification of these variables. In any cell that has seven or more observations, the largest multiple of seven that is less than or equal to the cell size are randomly assigned to the days of the week such that there are an equal number for each day of the week in this cell. For example, if there are 23 observations in a cell, a random 3 of them are assigned to each day of the week. The remaining 2 are not assigned in this step. Then the last variable in the list is dropped and 6-dimensional cells are created for the not-yet assigned observations. These are then assigned in the same way, and the process repeats until all variables have been dropped. The largest multiple of seven that is less than or equal to the remaining number of observations are again analogously assigned. In the first assignment on July 2, six observations remained, which were randomly assigned to Monday–Saturday. In the second assignment, five observations remained, which were randomly assigned to Sunday, Monday–Thursday, and so forth. So in subsequent assignments, we take into account which days had more assignments in the previous weeks, but we do not attempt to balance the assignments across weeks by the characteristics.

For the non-core sample, we follow the same procedure, with two differences. The first difference is that the cell may be missing because one or more of the variables used in the cross-classification may be missing. These observations are not assigned in this step. If the only missing was on the last variable, the cell in the next step will not be missing, and they may be assigned in that step. The second difference is that at the end of the process, we recode the days of the week in reverse order (Monday becomes Sunday, Tuesday becomes Saturday, etc.). This improves the balance of sample size across days of the week for the combined core and non-core sample.

The first invitations were available on Monday, July 4, 2016, and the first invitations for each newly assigned set of members are available the Monday after assignment. (For members who are assigned to a different day of the week, their invitations appear on their assigned day of the week after this Monday.)

3 Analysis sample

Each member is assigned to a day of the week. The member receives the invitation to complete the poll survey each week on the same day of the week. However, we allow them to respond up to six days later, that is, until the last day before they receive the next invitation.

We define a poll *week* as a period from Monday until the subsequent Sunday in which a member responded. So poll week 1 contains responses from July 4 up to and including July 10, poll week 2 contains responses from July 11 up to and including July 17, and so forth. Poll *waves* are defined in the same way, except that these use the date the invite was made available to define the wave, irrespective of the date of the response. Hence, the response of a member who was invited on July 7 but responded on July 12 is defined to be in wave 1 but week 2.

In the first 10 waves of the poll, 50% of the responses were on the assigned day. Another 23% were on the next day, and percentages decrease with the number of days until the last day (6 days after the invite), when the last 3% of responses are received.

In principle, each member is expected to respond once in each wave. However, as usual, not every member responds in each wave, so there is some nonresponse. The *completion rate* is the fraction of invitations that result in a completed survey. This was 69% in the first 10 waves (invitations up to September 4, responses up to September 10). There is some variation across waves, with completion rates varying between 66% in wave 10 and 72% in wave 6. Additionally, there are 121 observations in waves 2 and 3 by members who were scheduled to be invited for the first time in the next wave. We currently do not know how this could have happened. These are not included in the completion rate mentioned above, but we treat them as valid responses for the poll. Under some circumstances (for example, if they twice open a new browser window with the as-yet unfinished survey), it is technically possible that a member contributes more than one observation per invite. Up to September 16, it has happened 10 times that an invite led to two observations, and never more than two. Of these 10, three times the two responses were on the same day, with delays of 11 seconds, 2 minutes, and 41 minutes, respectively, between the two. In computing our poll results, we keep both responses from the same invite if they are on different days, but keep only the first response if they are on the same day.

Each day, a few minutes after midnight (Pacific Time), the data are downloaded and processed by a Stata program. The analysis sample for that day consists of (a subset of) the observations of the past seven days. For example, on Saturday, September 17, the observations from Saturday, September 10–Friday, September 16 are included in the analysis sample. For determining the date of a response in cases where a survey was started on one day and completed on the next, we use the date at the time of completion as the date of the response. We assign results for this seven-day period to the last day of the period (September 16 in the example) in the posted graphs on the website. Thus, we use a rolling window, and the results for two consecutive days are based on samples that overlap on average by 6/7th.

The analysis sample is then selected using the following rules (in this order):

1. If a member has provided multiple responses on the same day, keep only the first of these. As mentioned above, this has happened three times so far.

2. Drop observations that are inconsistent: percentages do not add up to 100 or chance of voting is outside the 0–100 range. The questionnaire software attempts to enforce consistency by showing a pop-up if the answers are inconsistent, but in rare cases (e.g., older browsers), we cannot enforce this. So far, this has happened only once, on August 4, where a respondent entered a 100% chance of voting for Trump, minus 50% for Clinton, and 50% for someone else.

3. Drop observations with missing weight variables. If one or more of the variables that are used in the weighting (gender, race/ethnicity, education, age, family income, household size, voting in 2012) are missing, the observation will not get a weight and therefore is excluded from the analysis sample. As discussed above, this same list was used to define the core and non-core sample in the assignment of the weekday. However, it is possible that members from the non-core sample become part of the analysis sample, if in the time between the assignment and the poll date they complete another "My Household" survey in which they provide the missing demographics. The demographics for each observation in the poll data are the latest available on the date the poll survey was completed. As of September 16, 2016, about 2.5% of the observations in the poll have one or more missings on the weighting variables.

4. Drop observations from the Native American and LA County subpanels. The LA County subpanel initially targeted families with young children, and the specific sampling strategy makes it difficult to compute appropriate weights. We are currently recruiting additional

members from LA County so that we will obtain a sample that can be viewed as representative of LA County as a whole. Once we have enough members from this additional batch, we should be able to construct weights that combine the LA County panel with the nationally representative panel and properly reflect population distributions, but currently, this is not possible yet. Hence, these do not obtain weights and are therefore excluded from the analysis sample. We do have a two-step procedure to compute proper weights for part the Native American subpanel, but this is only a small subset of the UAS (about 200 members), which is then heavily downweighted. Because of the additional complications caused by trying to implement this in an automated program and the small impact this would have on the poll results, we decided not to construct weights for this subpanel and drop their members from the analysis sample. Note that LA County residents and Native Americans in the nationally representative panel are included in the analysis sample, and receive proper weights.

4 Estimation

All analyses are weighted. See the weights document (`http://cesrusc.org/election/weights03.pdf`) for details about the construction of the weights. In the following, the weight variable for observation i in the analysis sample is denoted by w_i.

The most straightforward estimates are the ones for the respondents' subjective probability that each candidate will win the election. Let R_i and D_i be the probability that the Republican candidate (Trump) and the Democratic candidate (Clinton), respectively, will win the election as assessed in observation i. The graph on the "Respondents' predicted winner" tab of the Daybreak website plots the weighted averages of R_i and D_i:

$$\bar{R} = \frac{\sum_{i=1}^{n} w_i R_i}{\sum_{i=1}^{n} w_i} \qquad \text{(probability that Trump wins);} \qquad (1)$$

$$\bar{D} = \frac{\sum_{i=1}^{n} w_i D_i}{\sum_{i=1}^{n} w_i} \qquad \text{(probability that Clinton wins),} \qquad (2)$$

where n is the number of observations in the analysis sample. Note that this is computed separately for each analysis sample (each 7-day window), but we omit the subscript for that here.

Now consider our prediction of the popular vote. Each observation contains a subjective probability v_i that the individual will vote, a subjective probability r_i that the individual will vote Republican (Trump) if the individual will indeed vote, and analogously d_i for Democratic (Clinton). We also elicit voting for someone else, but we do not post the results for this. Let v_i be scaled such that it is between 0 and 1. Then the unconditional probability that the individual will vote Republican is $\rho_i = v_i r_i$ and similarly $\delta_i = v_i d_i$ for Democratic. The weights w_i are scaled to add up to population size, so the predicted numbers of Republican and Democratic votes are then

$$\hat{N}_R = \sum_{i=1}^{n} w_i v_i r_i = \sum_{i=1}^{n} w_i \rho_i \qquad \text{(votes for Trump)} \qquad (3)$$

and

$$\hat{N}_D = \sum_{i=1}^{n} w_i v_i d_i = \sum_{i=1}^{n} w_i \delta_i \qquad \text{(votes for Clinton),} \qquad (4)$$

respectively, and the predicted total number of votes is

$$\hat{N} = \sum_{i=1}^{n} w_i v_i \qquad \text{(total votes).} \qquad (5)$$

Hence, the predicted fractions of the popular vote for each of the candidates are

$$\hat{P}_R = \frac{\hat{N}_R}{\hat{N}} = \frac{\sum_{i=1}^{n} w_i \rho_i}{\sum_{i=1}^{n} w_i v_i} \qquad \text{(Trump's percentage of the popular vote)} \qquad (6)$$

and

$$\hat{P}_D = \frac{\hat{N}_D}{\hat{N}} = \frac{\sum_{i=1}^{n} w_i \delta_i}{\sum_{i=1}^{n} w_i v_i} \qquad \text{(Clinton's percentage of the popular vote).} \qquad (7)$$

These are ratios of estimated totals.

We also provide predictions of the popular vote and averages of the respondents' predicted winning chances by subgroup. Let g_i be a dummy variable that indicates whether the individual belongs to a subgroup of interest. For example, if the subgroup is "males", $g_i = 1$ for men and $g_i = 0$ for women. Then the subgroup estimates are computed in the same way as the estimates above, except that w_i is replaced by $w_i g_i$.

Finally, we provide estimates of voting intention by candidate preference. For candidate preference, we use the conditional probability of voting for each candidate, if the individual were to vote. For example, if an individual indicates a 70% chance of voting for Trump if he or she were to vote, and 30% chance of voting for Clinton, we count this individual as 0.7 Trump supporters and 0.3 Clinton supporters.

Hence, the estimated number of Trump supporters and Clinton supporters are

$$\hat{S}_R = \sum_{i=1}^{n} w_i r_i / 100 = \sum_{i=1}^{n} w_i \tilde{r}_i \qquad \text{(number of Trump supporters)} \qquad (8)$$

and

$$\hat{S}_D = \sum_{i=1}^{n} w_i d_i / 100 = \sum_{i=1}^{n} w_i \tilde{d}_i \qquad \text{(number of Clinton supporters),} \qquad (9)$$

respectively, with $\tilde{r}_i$ and $\tilde{d}_i$ implicitly defined. The predicted number of Republican and Democratic votes are $\hat{N}_R$ and $\hat{N}_D$ as discussed above. The estimated voting intentions of Trump supporters and Clinton supporters are the predicted percentages of Trump supporters and Clinton supporters that will vote. Consequently, these are

$$\hat{V}_R = \frac{\hat{N}_R}{\hat{S}_R} = \frac{\sum_{i=1}^n w_i \rho_i}{\sum_{i=1}^n w_i \tilde{r}_i} \qquad \text{(percentage of Trump supporters that will vote)} \qquad (10)$$

and

$$\hat{V}_D = \frac{\hat{N}_D}{\hat{S}_D} = \frac{\sum_{i=1}^n w_i \delta_i}{\sum_{i=1}^n w_i \tilde{d}_i} \qquad \text{(percentage of Clinton supporters that will vote).} \qquad (11)$$

These are again ratios of estimated totals.

5 Standard errors

As briefly indicated in the weighting document (`http://cesrusc.org/election/weights03.pdf`), we use a survey bootstrap to compute standard errors. Specifically, we declare the data to be survey data, with the individual as the primary sampling unit (cluster). This is because observations on the same individual are likely to be highly correlated.[1]

We construct 80 bootstrap replication weights, using a combination of the Stata user-written packages `survwgt` (Winter, 2015) and `bsweights` (Kolenikov, 2010) to compute 80 replication weights for each observation. This closely follows Example 6 in Kolenikov (2014), except that we use `survwgt` instead of `ipfraking` to construct the weights, because `ipfraking` was very slow and did not converge after 2000 iterations when we tried it for the first week's sample.

After constructing the main weight with `survwgt` as described in the weighting document, this procedure does the following steps 80 times:

1. Draw a sample of size $n - 1$ with replacement from the analysis sample.

2. Construct the weight for this sample in the same way the main weight was constructed. This gives each observation j of this bootstrap sample a weight w_j^*.

3. Observation j from the bootstrap sample is an observation (say i) from the original sample. An observation i from the original sample occurs k_i times in the bootstrap sample (k_i may be 0) and each time gets the same weight w_j^*. This information is encoded by giving i the replication weight $w_i^* = k_i w_j^*$ (zero if $k_i = 0$).

[1] One might argue that it would be even better to use the household as the cluster, or even the zip code the household was drawn from, for those households that were drawn as part of a two-stage sampling design. The poll does not currently implement this. A relatively small fraction of households currently has more than one member, so this will have a minor impact. The initial zip code is currently not easily available. However, the UAS's first and largest recruitment batch was a simple random sample from the database of our initial vendor ASDE, so these are not clustered.

Stata's svy facilities then use these weights to compute estimates of the covariance matrix of the estimators from section 4. This amounts to computing each estimator 81 times: once for the main weight, which gives the main estimate, and once using each replication weight. The estimated covariance matrix is the covariance matrix of the estimates from the 80 replication weights. See StataCorp (2011, p. 186).

The square roots of the diagonal elements of this estimated covariance matrix are the standard errors of the estimators from section 4. However, we are primarily interested in whether the difference between the estimate for Trump and the estimate for Clinton is significant and less so in the separate standard errors. Let $\hat{\theta}_R$ and $\hat{\theta}_D$ be any of the estimators from section 4 for Trump and Clinton, respectively. Denote by $\hat{V}_{RR}$ and $\hat{V}_{DD}$ the corresponding estimated variances and $\hat{V}_{RD}$ the corresponding estimated covariance. Then we are interested in the difference $\hat{\Delta} = \hat{\theta}_R - \hat{\theta}_D$. Its standard error is

$$\operatorname{se}(\hat{\Delta}) = \sqrt{\hat{V}_{RR} + \hat{V}_{DD} - 2\hat{V}_{RD}}. \tag{12}$$

The difference $\hat{\Delta}$ is statistically significant at the 5% level if $|\hat{\Delta}| > 1.96\,\operatorname{se}(\hat{\Delta})$.

6 Area of uncertainty

In the graphs, we would like to plot the individual curves for Trump and Clinton, but also indicate whether their difference is statistically significant. The individual confidence intervals are not suitable for this, because the two curves are highly (but not perfectly) dependent. For the predecessor of the Daybreak Poll, the RAND Continuous 2012 Presidential Election Poll, Kapteyn, Meijer, and Weerman (2012) developed the area of uncertainty for this (see also Gutsche, Kapteyn, Meijer, & Weerman, 2014). This is an area centered at the midpoint between the two curves whose width is $1.96\,\operatorname{se}(\hat{\Delta})$. If the curves are inside the area of uncertainty, their difference is not statistically significant, whereas if the curves are outside the area, the difference is significant.

Formally, the area of uncertainty is the area that for each analysis sample (represented by its last day) consists of the interval

$$\left[\frac{\hat{\theta}_R + \hat{\theta}_D}{2} - \frac{1.96}{2}\operatorname{se}(\hat{\Delta});\quad \frac{\hat{\theta}_R + \hat{\theta}_D}{2} + \frac{1.96}{2}\operatorname{se}(\hat{\Delta})\right]. \tag{13}$$

Suppose $\hat{\theta}_R > \hat{\theta}_D$, so that $\hat{\Delta} > 0$. (The reverse situation is analogous.) Then $\hat{\theta}_R$ is outside this interval if

$$\hat{\theta}_R > \frac{\hat{\theta}_R + \hat{\theta}_D}{2} + \frac{1.96}{2}\operatorname{se}(\hat{\Delta}),$$

which is equivalent to

$$\hat{\Delta} = \hat{\theta}_R - \hat{\theta}_D > 1.96\,\operatorname{se}(\hat{\Delta}).$$

When $\hat{\theta}_R$ is outside the interval (13), $\hat{\theta}_D$ is also outside this interval (on the other side).

7 Software implementation

As mentioned before, the analyses are done in Stata. This involves the following steps:

1. **processprior.do**
   Load the latest version of the UAS47 (pre-election survey) data, construct a variable that codes voting in the 2012 Presidential election, and save this processed data file.

2. **processpoll.do**
   Load the latest raw poll data, which includes for each observation the latest demographic information available at the time of the observation. Construct derived variables (e.g., poll date as a Stata date-format variable) and do some further cleaning and processing. Construct the variables that will be used in the weighting, except the prior voting variable. Check consistency of the answers and compute the unconditional probabilities of voting for each candidate. Save this processed data file.

3. **mkfulldata.do**
   Load and merge the two processed data files from the previous two steps. Create the prior voting variable, which combines the prior voting information from UAS47 with the age information from the demographics in the poll data (for the "too young" category). Save this in the file `fulldata.dta`, the latest version of which can be downloaded from the election pages by registered users of the UAS.

4. **mkpolldata.do**
   Load `fulldata.dta`. Keep only the analysis sample and the subset of the variables that is needed for the analyses posted on the website. Construct the weights. Save this in the file `polldata.dta`, the latest version of which can be downloaded from the election pages by registered users of the UAS. Note that the latest version of `fulldata.dta` should supersede all previous versions, but this is not the case for `polldata.dta`, which has weights that are specific to a certain 7-day window. However, it should be possible to recreate these at any time from the `fulldata.dta` file. We back these files up daily and will combine them and release this after the election.

5. **pollestim.do**
   Load `polldata.dta`. Construct a few variables that are used in the estimation. Compute the various estimates discussed in section 4, the Trump-Clinton differences associated with them, the standard errors of the Trump-Clinton differences, and the bounds of the area of uncertainty. Combine these results with results from prior days as stored in a number of small data files with the day-specific results. Save the combined results again, overwriting the previous versions of these results files. Export these results files to tab-delimited text files (with extension `.csv`). These text files can be downloaded from the election data pages, even by non-registered users. These files are used in

generating the graphs on the Daybreak Poll website of the Center for Economic and Social Research at USC (http://election.usc.edu), the graph on the website of the Jesse M. Unruh Institute of Politics at USC (http://dornsife.usc.edu/cf/unruh/poll.cfm), and the graphs on the Los Angeles Times website (http://graphics.latimes.com/usc-presidential-poll-dashboard/).

Some of the critical parts of the code are the following. Note that these contain variables and macros whose definitions are not obvious from this context. The goal here is to show the commands used, rather than the precise variable definitions. The complete Stata code is available upon request.

- Creating the main weight:

```
survwgt rake ‘basewgt’, by(‘vlist’) totvars(‘totlist’) generate(‘stub’main) ///
    maxrep(2000) check(1)
```

- Creating the replication weights:

```
svyset uasid [pw=‘basewgt’], strata(_one)
bsweights ‘stub’, reps(80) n(-1) seed(90089) ///
    calibrate(survwgt rake @, by(‘vlist’) totvars(‘totlist’) replace)
```

- Declaring the data as survey data, with the appropriate cluster, weight, and bootstrap replication weight definitions:

```
svyset uasid [pw=‘stub’main], vce(bootstrap) bsrweight(‘bsrlist’)
```

- Respondents’ predictions of chances of winning:

```
svy: mean trump_win clint_win other_win
lincom trump_win - clint_win
```

- Popular vote:

```
svy: ratio (Trump:   prob_trump/prob_vote1) ///
           (Clinton: prob_clint/prob_vote1) ///
           (Other:   prob_other/prob_vote1)
lincom Trump - Clinton
```

- Intention to vote by candidate preference:

```
svy: ratio (Trump:   prob_trump/trump_vote1) ///
           (Clinton: prob_clint/clint_vote1) ///
           (Other:   prob_other/other_vote1)
lincom Trump - Clinton
```

- Respondents’ predictions of chances of winning, by demographics:

  ```
  svy: mean trump_win clint_win, over(‘v’, nolabel)
  lincom [trump_win]‘val’ - [clint_win]‘val’
  ```

- Popular vote, by demographics:

  ```
  svy: ratio (Trump:   prob_trump/prob_vote1) ///
            (Clinton: prob_clint/prob_vote1), over(‘v’, nolabel)
  lincom [Trump]‘val’ - [Clinton]‘val’
  ```

References

Gutsche, T. L., Kapteyn, A., Meijer, E., & Weerman, B. (2014). The RAND Continuous 2012 Presidential Election Poll. *Public Opinion Quarterly*, *78*, 233–254. doi: 10.1093/poq/nfu009

Kapteyn, A., Meijer, E., & Weerman, B. (2012). *Methodology of the RAND Continuous 2012 Presidential Election Poll* (Working Paper No. WR-961). Santa Monica, CA: RAND Corporation. doi: 10.2139/ssrn.2146149

Kolenikov, S. (2010). Resampling variance estimation for complex survey data. *Stata Journal*, *10*, 165–199.

Kolenikov, S. (2014). Calibrating survey data using iterative proportional fitting (raking). *Stata Journal*, *14*, 22–59.

StataCorp. (2011). *Stata survey data reference manual: Release 12*. College Station, TX: StataCorp.

Winter, N. (2015). *SURVWGT: Stata module to create and manipulate survey weights.* Retrieved from http://EconPapers.repec.org/RePEc:boc:bocode:s427503