Classroom Probability Simulations Using R: Margin of Error in a Public Opinion Poll

Outstanding Professor Address, Part 2
Fall Faculty Convocation (9/21/04)

Bruce E. Trumbo
Department of Statistics, CSU Hayward
www.sci.csuhayward.edu/~btrumbo
“Margin of Error” in a Public Opinion Poll

Many polls quote a margin of sampling error — depending on the number $n$ of people sampled:

- $\pm 2\%$ for $n = 2500$
- $\pm 3\%$ for $n = 1100$
- $\pm 4\%$ for $n = 625$
For a candidate's lead: Double the margin of error.

The SF Chronicle for 9/18/04 printed a front-page report on apparent discrepancies of recent polls in the Bush vs. Kerry US Presidential race. For a poll with a 4% margin of error the report made a mistake in the margin of error for Bush over Kerry. Following is the text of my letter to the editor on this mistake:
Editor -- John Wildermuth's report ["Fickle polls give little away: Bush way ahead or just behind—depends on who asks", 9/18], while mainly excellent, contained a technical error about the margin of sampling error for the lead one candidate holds over another.

It is true that a poll based on 767 randomly chosen respondents, of whom 55 percent prefer Bush and 42 percent prefer Kerry, had a 4-point margin of sampling error. So the number for Kerry may well have been between 38 percent and 46 percent, and for Bush between 51 percent and 59 percent.

Bush's lead in this poll was 13 percent but, according to the laws of probability, the margin of sampling error for Bush's lead is twice as large -- about 8 points -- and his lead could have been anywhere between 5 points and 21 points.

The 4-point margin of sampling error can also be misleading when you compare two independent polls. Margins of sampling error in polls are usually based on the goal of being right 95 percent of the time.

When you compare a candidate's percentage (of likely votes) across two independent polls, the probability that both polls are correct drops to about 90 percent.

Taking all these issues into account along with others, the discrepancies among various polls are hardly surprising. Of course, it is possible that some polls are deliberately biased, but just the inherent inaccuracies in honest, state-of-the-art polling -- especially with samples smaller than 1,000 taken in a volatile political season -- are enough to account for the "fickle" performance of polls that fuel so much partisan commentary.

BRUCE E. TRUMBO / Professor of Statistics / California State University Hayward
What is the Basis for These Margins of Error?

- Assume a random sample.
- Ignore subjective issues.
- Based on probability rules: hard to prove. Don't like to say "Because I said so"?
- Simulation *illustrates* these rules.
- Requires faith that a computer can take a random sample: excellent software.
- We use R: State of the art — and free.
Simulating an Election Poll

To simplify, suppose:

- No undecided voters.
- 53% of voters favor Candidate “A”.
- 47% of voters favor Candidate “B”.
- Randomly sample 25 voters.

Can we detect from data on 25 subjects that Candidate “A” is in the lead?
One Simulation Using R:

\[
> \text{sample}(0:1, 25, \text{rep=T,} \\
\text{prob=c(.47, .53)})
\]

\[
[1] 1 0 1 1 0 0 1 0 0 1 0 0 0 0 0 1 0 1 0 1 1 1 0 1 1 0
\]

Interpret 0s & 1s as candidate preferences

\[
\begin{align*}
\end{align*}
\]

12/25 = 48% for “A”.
What can we conclude from this survey of 25 subjects?

Poll result after 25: “A” has 12 in favor: Proportion = 48% for “A”.

Contradicts what we know about population. Intuitively, we know 25 is not a large enough number.

How to look at this more objectively?
Find Proportions After Each Subject

Table Below Used to Find 1st Through 5th Proportions

<table>
<thead>
<tr>
<th>Subj. so Far</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opinion</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A’s so Far</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Prop. so Far</td>
<td>1/1 = 1.00</td>
<td>1/2 = 0.50</td>
<td>2/3 = 0.67</td>
<td>3/4 = 0.75</td>
<td>3/5 = 0.60</td>
</tr>
</tbody>
</table>

Then plot a trace of the poll.
Now plot the remaining 20 subjects.
Endpoint is 48%. What would other polls say?
Unstable Trace of a 25-Subject Poll

Note: Here and below, vertical axis from 25% to 75% to help show detail.
Unstable Trace of a 25-Subject Poll

Proportion Favoring 'A'

Number of Subjects
Unstable Trace of a 25-Subject Poll

Proportion Favoring 'A'

Number of Subjects
Results Differ Widely. On one graph, now overlay 20 traces.
All traces **end** between 40% and 70%, most above 50%.
Vast majority of traces end between 30% and 75%. Misleading: 377 of the 1000 traces end below 50%.
Results fit a predictable, but useless pattern. A 25-subject poll can’t show if “A” is winning.
Polls with $n = 2500$ Subjects

Major polling organizations report such polls as having a ±2% margin of sampling error.

**Intended Meaning:**
If the true proportion for “A” is 53%, then most 2500-subject polls (95%) will have endpoints between 51% and 55%, thus detecting that “A” is the favorite.
Ultimately Stable Trace of a 2500-Subject Poll

Proportion Favoring 'A'

Number of Subjects
Ultimately Stable Trace of a 2500-Subject Poll
Ultimately Stable Trace of a 2500-Subject Poll
Ultimately Stable Trace of a 2500-Subject Poll

Proportion Favoring 'A'

Number of Subjects
Red marks at right indicate 53% ± 2%. Green line at 50%.
On “average” run: 19 out of 20 traces (95%) within margin of error.
95% of endpoints within margin of error; only 3 in 1000 below 50%.
In the interval from 51% to 55% for Candidate “A”:
Black curve has 95% probability. Blue curve only about 12%.
Not all processes settle to a meaningful single limiting value, such as we have seen for polls. Here are 3 examples.

The first shows a kind of Brownian motion process (uniformly distributed displacements around a circle); it approaches no limit.

The second shows a brother-sister mating process (relax, it's for plants); it can approach either of two limits.

The third shows what can happen with a random number generator of poor quality; it converges to an incorrect limit. (This illustrates the importance of using quality software such as R.)
A Process With No Target Value

From an example in a forthcoming book by E. Suess and B. Trumbo (Springer).
From an example in a forthcoming book by E. Suess and B. Trumbo (Springer).
A bad pseudorandom number generator (black) causes rapid convergence to an incorrect limit. A good generator (green), slower convergence to the correct limit.
Some processes approach limiting distributions rather than limiting values. The limiting distribution is sometimes, but not always normal. Three examples follow: The first shows results from 5000 experiments, each tossing 10,000 fair coins. It is relatively unlikely for “Heads” to be in the lead anywhere near half the time.

The second plots sample means against sample standard deviations for samples of size 5 from a bathtub-shaped beta distribution. The relationship between means and standard deviations is intricate.

The third (mainly recreational) example shows that it is possible for the limiting distribution of a process to look like a nonmathematical object.

[The "bathtub shaped" distribution is BETA(1/2, 1/2).]

Image description:
- A scatter plot with the title "BETA(.2,.2): 20000 Samples, n=5".
- The x-axis represents the sample standard deviation and the y-axis represents the sample mean.
- The plot shows a distribution of data points.
Simulation Papers Presented by CSU Hayward Students at Joint Statistics Meetings, Toronto, Canada, August 2004.

Papers illustrate pedagogical uses of R. Our sessions were extraordinarily well attended because of the current interest in using simulation to teach statistics.

There are $n$ people in a room. What is the probability two (or more) have the same birthday?

If birthdays are uniform throughout the year, then the following graph (based on a simple formula, Stat 3401) gives the answer.
There is better than a 50-50 chance of a match when more than 23 people are in the room. Almost sure when there are more than 50.

But US birthdays aren’t uniform (36 months shown): 104% of average in Summer; 96% in Winter—except for a spike of “tax deduction babies” in December.
Simulation shows actual US nonuniformity doesn’t much change the probability of getting “no match,” but extreme nonuniformity (left side of each graph) would.

Assume half of annual births uniformly in “Weeks” and the other half uniformly in the rest of the year. “Weeks=26” is uniform over the whole year.

Each point above is based on 50,000 simulated rooms, with 25 people in each room.
B. Trumbo, E. Suess, R. Brafman: “Classroom Simulation: Are Variance-Stabilizing Transformations Really Useful?”

If data are exponentially distributed (instead of normally), some statistical procedures, such as an analysis of variance, may not work as they should. (Stat 3503.)

Simulation shows when log and rank transformation of such data are helpful.
Power Against Various Alternatives

When $M = 1$, $H_0$ is true; when $M = 2$, the group means are 1, 2, 4; and when $M = 4$, the group means are 1, 4, 16; etc.

Simulations for this graph required generating about 3 million ANOVAs, each with 15 observations.

These transformations are helpful when samples are small.