Kaplan Chap 3 / Example 10: The Statistical Power of Polling

The Presidential election has come and gone. Anderson got 9% of the vote. He didn't qualify for federal campaign support in the next election, but he did split the right-wing vote.

On Public Television, Bill Moyers is hosting a round-table discussion about the media coverage of the campaign. The issue is Anderson's reported “surge" in support in the last days of the campaign. (See Example??.) As facts turned out, the surge never materialized. Reporters have egg on their faces.

A statistician is critical of the reporters: “There was never any reason to report a surge. The p-value of the reported surge was 11 percent -- no reason to reject the Null Hypothesis of no change in support."

Cynthia Brokow retorts, “That's for a two-sided test. The one-sided p-value was about 5 percent. It's our responsibility to keep the public informed and not to suppress information because it doesn't reach some ivory-tower threshold for reliability. We gave the raw numbers from the poll; it's up to the viewer to figure this out."

Moyers mediates. “Clearly there's a problem here. We reported a story that was wrong and for which, in hindsight, we didn't have much evidence. If we reporters can't digest these statistics, how can we expect the public to do so? There must be some balance between reporting the raw facts and reporting only those facts which, with due statistical consideration, provide a reasonable level of support for the conclusions they seem to point to."

Steven Brill, editor of a media watchdog magazine, has a suggestion. “This is a question of standards and responsible reporting. We have an obligation to collect enough data to make our results reliable, particularly when the results are important. The problem is in the size of the poll. The polls have to be big enough so that we when claim something as remarkable as a 3 percent increase in support we have good reason to believe the data. I don't care if we make mistakes with claims concerning 1 percent changes in support, but we have to be right when claiming 3 percent."

Moyers: “Well, how big does such a poll have to be?" All eyes turn toward the statistician.

The statistician: “This is an example of a sample-size calculation. We want to make the sample size large enough so that our hypothesis test has a low *signifcance level* against the null hypothesis and a high *power* against the alternative hypothesis. As you know, there's generally a trade-off between power and significance, and ..."

Moyers interrupts. “Hold on a second. Let's bring this down to Earth. I don't know much about statistics but as a reporter I know that we want our stories to have a high significance level."

The statistician: “Sorry. I was using some technical terms whose meaning doesn't always correspond well to the everyday meaning of these words. The *null hypothesis* is a statement which we are going to *reject* or *not reject* on the basis of our data."

Reporter: “Like, `Nothing much has happened. No change in support."

Statistician: “Exactly. The null hypothesis plays the role of the devil's advocate. We also have a *test statistic* | in this case that's the fraction of support measured in our poll, or, rather, the change in the fraction of support between the two polls. And, we have a *rejection threshold* that measures what we're interested in. This is a level we set ahead of time. If our test statistic is beyond the threshold level then we conclude that the data justifies rejection of the null hypothesis."

Reporter: “What about the p-value?"

Statistician: “The p-value is something we calculate after we already have our data. Right now we're discussing how to design the poll, not how to analyze the data from the poll."

Moyers: “So where does the sample size come in?"

Statistician: “The sample size determines where we set our rejection threshold so that our conclusions are reliable. Imagine that the devil's advocate is right and the null hypothesis is true. Since we're randomly picking the voters questioned in the poll, it might happen that our poll results are above the rejection threshold just by chance. So, even if the null hypothesis is right our poll results might cause us to reject the null."

Sam Donaldson: “That's pretty unlikely."

Statistician: “I don't know how you can say that since we haven't yet set either the threshold or the sample size. Of course, you're right in the sense that our goal here is to set the sample size and threshold so that the probability of falsely rejecting the null hypothesis is very small. A false rejection of the null is called a *Type I error* and the probability of making such an error for a given null hypothesis, threshold, test statistic, and sample size is called the signficance level of the test. We want a low significance level, that is, a low chance of making a Type I error."

Moyers: “How do we find the significance level?"

Statistician: “Since the sample size is what we want to figure out, first we pick a rejection threshold."

Moyers: “OK. How do we pick a rejection threshold?"

Statistician: “Bear with me | we'll come to that in a bit. For now, let's assume that you already have the threshold, say a change in the polls of 1%. This means that if the test statistic -- which is the difference between successive polls -- is more than 1%, we will reject the null hypothesis. We want to make the sample size large enough to make the significance level small."

Moyers: “How small a significance level is small enough?"

Reporter from the McLaughlin Group: “50%"

Reporter from CBS news: “25%. We have high standards."

Reporter from the Christian Science Monitor: “0%. We want to be right all the time."

Statistician: “The standard in scientific research is 5%. This means that if the null hypothesis is true, we'll make a mistake about one time in 20. If we set the significance level at 50%, even when nothing is happening we will have a headline story -- albeit wrong -- for just about every second poll."

National Inquirer: “Exactly. That's the nature of our business. The public has a right to know."

Statistician: “On the other hand, it would be a mistake to insist that we never make a Type I error for example insisting on a 0% significance level. Doing so would practically ensure that we would always make Type II errors."

Moyers: “Type II errors?"

Statistician: “A *Type II error* occurs when the null hypothesis is wrong, but we fail to reject the null."

Moyers: “That would happen, for instance, if there were a big change in support from one poll to the next but our rejection criteria were so rigorous that we refused to conclude that something had changed."

Statistician: “Right. What we want to do is set our rejection threshold to make both types of error unlikely. Unfortunately, there is a tradeoff between making the two types of error. For instance, we can lower the probability of a Type I error by making the rejection threshold harder to satisfy."

Moyers: “You mean by saying that we won't report that there has been a change in support unless the difference from one poll to the next is at least 2%, not 1% as previously suggested."

CBS news: “But that would make it less likely that we'd be able to report a change."

Statistician: “Right. That would be good, though, if there really were no change. You'd avoid a Type I error."

CBS news: “But what if there were really a change in support?"

Statistician: “Then not reporting it would be a Type II error. As I said, there's a trade-off between Type I and Type II errors. If you alter the rejection threshold to reduce the probability of making one type of error, you increase the probability of the other type."

Moyers: “Fascinating. But where does the sample size come in?"

Statistician: “There is one way around the trade-off. We can reduce the probabilities of both types of error by making the sample size large."

Moyers: “How large?"

Statistician: “The larger the better. But in order to make the poll economically feasible, you also want to make the sample size small. So, I'll calculate the minimum acceptable size of the sample. First, I need to compute the probability of a Type I error. What's your null hypothesis?"

Sam Donaldson: “That there has been no actual change in the level of support."

Nina Totenberg: “But what will we be justified in reporting if we reject the null; only that `support has increased.' That's not a very strong statement."

Statistician: “Right. Perhaps you'd rather have a stronger statement. If your null were `support has changed by less than 1%' then if you reject the null you'll be able to make a stronger statement."

Moyers: “Let's take `less than 1%' as our null."

Statistician: “We'll use a significance level of 10% for the calculation. Now ... What's your alternative hypothesis?"

Moyers: “You mentioned that at the beginning. What is that?"

Statistician: “The *alternative hypothesis* is something that, if true, would lead you to reject the null."

Moyers: “Why not just take the alternative to be that the change in political support was greater than 1%. That's what we know if the null isn't true."

Statistician: “Good point. However, I need a specific hypothesis so that I can calculate the probability of a Type II error. Is it alright if I say that the alternative is, 'The real change in support was 3%?' "

Tottenberg: “Why not say 2.1%?"

Statistician: “We could. But before deciding, let's pick an acceptable error rate for Type II errors. If there really was a change of 3%, how much chance are you willing to take that you make a mistake and fail to
reject the null?"

Moyers: “That's difficult to answer. Failing to report something doesn't seem like as serious an error as reporting something that is wrong. Let's say that we're willing to miss the story 25% of the time."

Statistician: “OK. I have the information I need. By the way, the *power* of the hypothesis test is 1 minus the probability of a Type II error. That's 75% in this case and is the probability that we do (correctly) reject the null when the alternative is true."

I'll do the calculations for a case that's like the Ross Anderson situation where the background level of support is about 10%."

The statistician writes the following program using MATLAB and the Statistics Toolbox. It can be found in a script file named ‘**pollsize.m’**.

```
function [nullthreshold, beta] = pollsize(sampleSize)
%'nullthreshold' is the critical poll value needed to reject the null
%'beta' is the probability of making a Type II error

seedRNG0;  %seed random number generator

backgroundsupport = 10;  %assumed percentage points of support for candidate
nullincrease = 1;        %minimum number of points increase considered
meaningful
altincrease = 3;         %criterion point improvement being searched for
alpha = 0.10;            %type I error setting

nIterations = 1000;
stats = zeros(nIterations,1);
%universe representing background support for candidate
electorate = [zeros(1,(100-backgroundsupport)), ones(1,(backgroundsupport))];
%universe representing background support plus nullincrease threshold
electorateplus = [zeros(1,(100-backgroundsupport-nullincrease)),
ones(1,(backgroundsupport+nullincrease))];
%universe representing background support plus alternativeincrease
%(altternative hypothesis)
electoratealt = [zeros(1,(100-backgroundsupport-altincrease)),
ones(1,(backgroundsupport+altincrease))];

for i=1:nIterations
    %simulate the difference between two polls assuming null hypothesis
    %plus 'nullincrease' threshold requirement
    poll1 = randsample(electorate, sampleSize, true);
    poll2 = randsample(electorateplus, sampleSize, true);
    %compute difference between random samples
    stats(i) = (sum(poll2)-sum(poll1))/sampleSize;
end

%find the critical value (threshold) needed to reject the null hypothesis
%(one-tailed test) using the cumulative bootstrap distribution
[cumprobs, xvals] = ecdf(stats);
%now, find the location of the targeted cum probability in the distribution
temp = abs(cumprobs - (1-alpha)); %subtract %ile to produce minimum
index = find(temp == min(temp));  %find location of minimum difference
nullthreshold = xvals(index);  %extract critical value needed to reject null

%Using the nullthreshold, compute the type II error rate under the
%alternative hypothesis scenario
for i=1:nIterations
    %draw sample from null ditribution
    poll1 = randsample(electorate, sampleSize, true);
    %draw sample from alternative hypothesis distribution
    poll2 = randsample(electoratealt, sampleSize, true);
    %compute difference between random samples
    stats2(i) = (sum(poll2)-sum(poll1))/sampleSize;
end
```

(script continues on next page)

```
%search alternative resampling distribution to find cumulative probability
%delineated by the null hypothesis critical value (threshold)
[cumprobs, xvals] = ecdf(stats2);
%now, find the location of the effect in the distribution
temp = abs(xvals - nullthreshold); %subtract %ile to produce minimum
index = find(temp == min(temp));  %find location of minimum difference
beta = cumprobs(index);  %extract cum prob of experimental effect
```

This program will take any sample size and compute the rejection threshold and the Type II error rate. It's assumed that the significance level (alpha) is 10%. We try this out for many sample sizes until we find the smallest one that gives us a reasonable Type II error rate. Then we just read off the appropriate rejection threshold.

Statistician: “Let's try a sample size of 100 (in each poll).
[threshold, errorrate] = pollsize(100)
ans: 0.06 0.70
We get a threshold of 6% and a Type II error rate of 70%. "

Sam Donaldson: “You mean that we won't say that there has been a change in support unless the polls have changed by 6%. That's ridiculous."

Statistician: “I agree. It means that the sample size is too small. Let's try 500.
[threshold, errorrate] = pollsize(500)
ans: 0.034 0.570
Now the threshold is 3.4% and the Type II error rate is 57%. This is still much too high. So let's try a much larger sample size.
[threshold, errorrate] = pollsize(2000)
ans: 0.023 0.230
Good. The Type II error rate is down to 23%, close to the specified value (of 25%).
The threshold is 2.3%. That seems to fit the bill, but barely. So, this is the smallest sample size that's acceptable."

Moyers: “I notice that you didn't really use our null hypothesis that support changed by less than 1%. Instead, you assumed that support changed by exactly 1%. Why?"

Statistician: “I wanted to make the computations conservative, so I took the worst possible case."

Moyers: “But how do you know this isn't too conservative. It's awfully expensive to poll 2000 people."

Statistician: “In order to do the calculation differently, I would need some more information: under your null hypothesis how likely is it that the real change is 0%, 1%, and so on. I don't see how you can possibly know this. But, if you think you do, you might want to contact a Bayesian statistician, or read Chapter 4 of Kaplan."

Moyers: “Let's summarize. We now have some standards for this particular case where we take two polls and want to say whether there is been a change in support for one candidate.

We should use random polls with at least 2000 voters. If the change in support level is greater than a threshold of 2.3% we are justified in reporting our results as indicating a change in support greater than 1%."

Donaldson: “But what if the measured change in support were greater than 10%. I'd feel pretty silly reporting only that the change is greater than 1%."

Statistician: “True. In fact, you could always make another null hypothesis |--say, the support change is greater than 8% -- and compute a p-value for your data against that null. If the p-value is low enough, you'd be justified in reporting that the change is greater than 8%. Remember, the null and alternative hypotheses here were framed for the purpose of figuring out how many people to interview in the poll. Once you have the data in hand, these hypotheses are of no particular relevance."

After a pause, the statistician adds: “Please remember that these results apply only to an increase in support for the underdog. If you want to report either an increase or a decrease, we need to do a two-tailed calculation and the sample size would need to be bigger."

Notes on origin of this exercise:
Text is from Chapter 3 of online text “Resampling Stats in MATLAB” by Daniel T. Kaplan
URL: http://www.resample.com/support/user-guides/resampling-stats-for-matlab-users-guide/
The MATLAB script ‘pollsize’ has been completely rewritten so that it does not require the Resampling Stats library (FS 6 January 2014)