

Lab 4: Re-randomization of randomized controlled trials (II)

Due 2/3, 3 Questions

[Note: This lab is a little, um, chatty. The actual computations are a bit light. For those of you missing messy data, we'll have plenty of it next week! This lab is a bit more thoughtful.]

We will finish our examination of randomized controlled trials with a comment of sorts about statistical significance. As you will recall from lecture, the painkiller Vioxx was pulled from the market in 2004 by its manufacturer, Merck. For several years prior, the scientific community had expressed concerns that long-term use of the drug increased a patient's risk of heart attack and stroke. Ultimately, Merck conceded the point, when one of their own studies (the so-called APPROVe trial or Adenomatous Polyp PRevention On Vioxx) found that patients on Vioxx were almost twice as likely to experience heart problems than those taking naproxen (Aleve). The APPROVe trial was ended early when this trend became apparent. You can read about that trial (complete with enough information to make your own 2x2 table and test the results) at <http://www.ncbi.nlm.nih.gov/pubmed/15713943>.

Our lab today will not deal with this study, but instead, one that took place a couple years earlier. As we mentioned in class, Vioxx was a very successful, high-profile drug; some 80 million patients took Vioxx, and Merck earned over \$2 billion dollars from the drug in 2003 alone. Given its popularity, there were many many studies examining its effectiveness in treating a variety of conditions (in class, for example, we talked about trials that examined its use for patients with osteoarthritis). Today we will consider a the so-called ADVANTAGE trial (Assessment of Differences between Vioxx and Naproxen To Ascertain Gastrointestinal Tolerability and Effectiveness).

The study appeared in the Annals of Internal Medicine in 2003 (although it was completed in 2001; evidently it was submitted to several journals and rejected because the results were "redundant" given previous trials). Below is a link to the paper.

http://www.vioxxdocuments.com/Documents/AIM/Lisse_AIM_Advantage_2003.pdf

The paper is a good example of studies of this kind, if only because you can see how the randomization works (page 542). Also notice that there is no mention of heart attacks in the summary sections of the paper. For a report appearing in 2003, it seems odd, given the track record that Vioxx had, that there would be no comment at all. If you skim down a bit you will find on page 543 the following sentence “Five myocardial infarctions occurred in the rofecoxib group, and 1 occurred in the naproxen group ($P>0.2$).” The P here is a P-value, of the sort we computed in class. (A myocardial infarction is a heart attack.)

1. Getting started

We will first load up a couple functions that we’ll need for this lab.

```
source("http://www.stat.ucla.edu/~cocteau/stat13/data/advantage.R")  
  
ls()
```

You should see a function called `theoretical` and a data set called `advantage`; we’ll get to these shortly.

We are going to start by recreating the statistical results from the ADVANTAGE trial. In all, there were 5557 patients recruited, with 2772 receiving naproxen and 2785 receiving rofecoxib or Vioxx. There were 5 cardiac events in the Vioxx group and just 1 in the naproxen group. In table form, this is

	N	V
no MI	2771	2780
MI	1	5

In the previous lab, we simulated tables under the null hypothesis that the treatment and control posed the same health risks (in this case, it would be that naproxen and Vioxx pose the same risk of heart attack). We saw that we could focus on one element of the table, or, in this case, the number of patients who had a heart attack under Vioxx. The data set `advantage` (type in its name and hit enter to have a look) is similar to `cannon` from the last lab. It has 5557 rows, one for each patient; the variable `outcome` describes whether or not they had a heart attack (MI or no

MI) and `treatment` says whether they received Vioxx or naproxen (V or N). The command

```
results = table(advantage)

results
```

should give you the table above. If you want, you can repeat the simulation exercise from last time using `advantage` instead of `cannon` to generate an estimate of the null distribution we're interested in.

In lecture, however, we worked out the exact distribution for this number under the null hypothesis, and we review the calculations in the appendix for this particular experiment. You can compute the distribution as follows

```
theoretical(results)
```

By default, the function will take your two-by-two table and generate the null distribution for the counts in the lower right hand corner (in this case, re-randomizing and seeing how many patients have heart attacks under Vioxx). You should see two columns, one for the count in the lower right hand corner of the re-randomized tables and the other for the probability of seeing that count. You can capture this data into a variable and make a plot with the following commands.

```
null = theoretical(results)

plot(null$k,null$p,type="h")
```

The argument `type` tells R that you don't want a scatter plot or a line plot, but instead a plot of "heights".

In lecture, we had been considering mainly one-sided tests; in this case, it means looking for evidence against the null from re-randomized tables that had at least as many heart attacks in the Vioxx group as our observed table. The term "one-sided" implies that we look for extremes in just one of the tails of the null distribution. Here, that would involve computing a P-value from the chance of seeing just 5 or 6 deaths under re-randomization.

Question 1: Compute the conditional proportions of people who have heart attacks under naproxen and Vioxx. What do you think about these numbers? Now, compute

the one-sided P-value associated with the null hypothesis of no difference between Vioxx and naproxen. If we used a P-value threshold that declared as statistically significant values below 0.05, would this result be “significant”?

As we mentioned in lecture, a one-sided test makes sense in this context because we know that naproxen does not increase a patient’s risk of heart attack. It is, however, commonplace to run this test “two-sided” and that is the specification given by the authors in their paper (search the paper for “Fisher’s exact test”). A two-sided test means that we look to both tails of the distribution for evidence against the null. Forming a P-value this way implies that we are open to the fact that naproxen and Vioxx might be different, but that a difference could mean either Vioxx causes more deaths than naproxen or vice versa. In that case, we take as “extreme” values in then null distribution, `null`, that have probability as small or smaller than the value associated with our data.

So, looking at the plot of `null` and comparing the heights of 0, 1,... 5, and 6 to the height for 5 (the probability that a re-randomized table assigns 5 heart attacks to the Vioxx group, 5 being the value we observed in the ADVANTAGE study), we see that 0, 1, 5 and 6 all qualify as being “as extreme or more extreme” than our observed data, the value 5.

In the last lecture, we noted that the difference between a one- and a two-sided test could be captured in our alternative hypothesis. That is, to run a one-sided test here, we have null hypothesis that Vioxx and naproxen pose the same risk of heart attack, with an alternative that Vioxx has a greater risk. In that case, we seek evidence in the upper tail of our null distribution; places where re-randomization assigns at least as many heart attacks to the Vioxx group as we observed in the data.

In a two-sided test, our null is again that Vioxx and naproxen have the same risk, the alternative is simply that they are not the same. Here, we have to look for evidence against the null by considering excess numbers of heart attacks in either the Vioxx group or the naproxen group.

As we mentioned above, to perform the two-sided test, we’d add up the probabilities for 0, 1, 5 and 6. Rather than torture you with that addition, we’ll start to make use of R as a *statistical* tool!

```
rerandomize.test(results)
```

This will perform the addition you need and report the two-sided P-value.

Question 2: If we use a 0.05 cutoff for statistical significance, are the 5 heart attacks under Vioxx (compared to the 1 in the naproxen group) significant? How does the value of this P-value compare to the one in Question 1? Bigger? Smaller? Why?

2. Statistical significance

We now come to a commentary on statistical significance. Recall that the 5 deaths in the Vioxx group are not commented on in the ADVANTAGE paper. The fact that the deaths were not sufficiently large in number to be considered significant meant that the authors could be somewhat quiet about the subject (although you could argue that given all the press about Vioxx, they had some responsibility to talk about these 5). Some time after the study appeared in the Annals of Internal Medicine, this story broke: <http://www.nytimes.com/2005/04/24/business/24drug.html>. It turns out that Merck had neglected to report 3 extra heart attacks in the Vioxx group. Do these three heart attacks change things?

Suppose just one extra heart attack was found, what would change? You can answer this question by changing one of the patients outcomes from “no MI” to “MI”. So, for example, look at patient 6

```
advantage[6,]
```

For publication, the researchers thought this patient did not have a heart attack. If this was one of the newly discovered heart attack victims, you can change their status with the command

```
advantage$outcome[6] = "MI"
```

Now, remake your `results` table and rerun the `rerandomize.test` and see what happens.

Question 3: Change the status of two more patients as we did above. Do this one patient at a time. With each, repeat your analysis in the first part of the lab, making plots of the null distribution. What do you notice? Next, consider the re-

randomization tests; how do the results of `rerandomize.test` change? Finally, if we were to apply a threshold of 0.05 to declare results as statistically significant, what do you find? What does this say about using strict thresholds blindly?

The point of this second part of the lab is that there are circumstances when a hard threshold feels inappropriate. Given the press around the issues related to Vioxx prior to this study (remember, this was published in 2003 and questions about Vioxx emerged years earlier), some sort of caution was warranted. The story only gets worse when we read that Merck knew about the extra heart attacks, incidents that would have, if properly tallied, been declared significant. We shouldn't speculate on whether any of this was intentional, but instead wonder about a system that forces a hard line, that incentivises researchers to be on one side or the other.

Appendix

Here we will work out the theoretical distribution of the number of heart attacks in the Vioxx group under re-randomization. We will use the original data set (the table at the beginning of the lab); obviously, the numbers will change if we change a patient's status. Therefore, we have 5557 patients total, 6 experienced a heart attack, 5551 did not. Under the null, the 6 patients who had a heart attack would have done so no matter what group they were randomized into. During our re-randomization, we could see 0, 1, 2, ... 5, or 6 of these patients in the Vioxx group. The probability of seeing, say, k patients who had heart attacks assigned randomly to the Vioxx group is given by

$$P(k \text{ heart attacks in the Vioxx group}) = \frac{\binom{6}{k} \binom{5551}{2785-k}}{\binom{5557}{2785}}$$

The downstairs in this fraction is the total number of ways we can select from the 5557 patients a group of 2785 to receive Vioxx (the remaining 2772 getting naproxen). The upstairs consists of two parts: The first is the number of ways we can select a group of size k from the 6 people who died of a heart attack; and the second is the number of ways we can select the remaining $2785-k$ patients (to fill out the Vioxx group) from the 5551 who did not suffer a heart attack during the trial.

To derive this equation we used the fact that each randomization, each division into treatment and control, each group we select are all equally likely. There are 5557-choose-2785 ways of making the split, so each has probability $1/(5557\text{-choose-}2785)$. To find the chance that a re-randomization will put exactly k people who experienced a heart attack into the Vioxx group, we just have count the number of randomizations that satisfy this condition. That number, as a proportion of the total 5557-choose-2785 different treatment-control assignments, is the probability we want.

This expression above is also known as the Hypergeometric distribution (and values can be obtained from the R command `dhyper`). It describes the process of “sampling without replacement.” In classical probability, you will find it describing draws from an urn filled with some number of black and white balls. By analogy, we could imagine an urn with 5557 balls, 6 white and 5551 black. What we are computing above is the chance that when selecting 2785 balls from the urn, we have just k white balls, where k is between 0 and 6.