Sharing data: Here's how you do it, and here's how you don't

Posted by Andrew on 15 September 2016, 9:44 pm

I received the following email today:

> Professor Gelman,
>
> My name is **, I am a senior at the University of ** studying **, and recently came across your paper, "What is the Probability That Your Vote Will Make a Difference?" in my Public Choice class. I am wondering if you are able to send me the actual probabilities that you calculated for all of the states, as some are mentioned in the paper, but I can't find the actual data anywhere online.
>
> The reason I ask is that I am trying to do some analysis on rational voter absenteeism. Specifically I want to see if there is any correlation between the probability that someone's vote will make a difference (From your paper) and the voter turnout in each state in the 2008 election.
>
> Thanks!

Hmmm, where are the data? I went to the page of my published papers, searched on "What is the probability" and found the file, which was called probdecisive2.pdf, then searched on my computer for that file name, found the directory, came across two seemingly relevant files, electionnight.R and nate.R, and send this student a quick email with those two R files and all the data files that were referenced there. No big deal, it took about 5 minutes.

And then I was reminded of this item that Malte Elson pointed me to the other day, a GoFundMe website that begins:

> My name is Chris Ferguson, I am a psychology professor at Stetson University in DeLand, FL. In my research, I'm studying how media affect children and young adults.
>
> Earlier this year, another researcher from Brigham Young University published a three-year longitudinal study between viewing relational aggression on TV and aggressive behavior in the journal Developmental Psychology. Longitudinal studies are rare in my field, so I was very excited to see this study, and eager to take a look at the data myself to check up on some of the analyses reported by the authors.
>
> So I spoke with the Flourishing Families project staff who manage the dataset from which the study was published and which was authored by one of their scholars. They agreed to send the data file, but require I cover the expenses for the data file preparation ($300/hour, $450 in total; you can see the invoice here). Because I consider data sharing a courtesy among

> researchers, I contacted BYU's Office of Research and Creative Activities and they confirmed that charging a fee for a scholarly data request is consistent with their policy.
>
> Given I have no outside funding, I might not be able to afford the dataset of Dr. [Sarah] Coyne's study, although it is very important for my own research. Although somewhat unconventional, I am hoping that this fundraising site will help me cover parts of the cost!

The paper in question was published in the journal Developmental Psychology. On the plus side, no public funding seems to have been involved, so I guess I can't say that these data were collected with your tax dollars. If BYU wants to charge $300/hr for a service that I provide for free, they can go for it.

Here's the invoice:

In future perhaps journals will require all data to be posted as a condition of publication and then this sort of thing won't happen anymore.

P.S. Related silliness here.

Filed under Sociology | Permalink

32 Comments

1. *Leo Kenji* says:
   September 15, 2016 at 11:32 pm

   from the wikipedia: "An experiment is a procedure carried out to support, refute, or validate a hypothesis. Experiments provide insight into cause-and-effect by demonstrating what outcome occurs when a particular factor is manipulated. Experiments vary greatly in goal and scale, but always rely on repeatable procedure and logical analysis of the results. There also exist natural experimental studies."

   If we're talking about science, a paper, IMO, only makes sense if it contains enough information to reproduce the results given the same circumstances ("repeatable procedure"). Conclusions and methodology without the used data, IMO, is not science.

   So I agree with you 100%.

   - *Carlos Ungil* says:
     September 16, 2016 at 1:20 am

     At least they are ready to provide the data for a (hefty) fee. Sometimes the response is more like "Why should I make the data available to you, when your aim is to try and find something wrong with it."

     - *Keith O'Rourke* says:
       September 16, 2016 at 8:23 am

Sometime they send you a non-disclosure agreement to sign first or simply please direct any further inquiries to our lawyers.

2. *Llewelyn Richards-Ward* says:

September 15, 2016 at 11:54 pm

What a brilliant idea charging for data? I wonder how much those darned participants billed the university to participate in those studies. I think it is just great that they are trying to ensure that information that benefits their careers and which they gather, one imagines for the betterment of humankind, knowledge and possibly world peace, also turns a profit. It has become quite concerning in social research that there is a trend toward open source data and replicability. This could lead us all down some dangerous paths. What if the governments, for example, allowed us to collect data we provide as part of our citizenship? All manner of things, up to and including making public servants and policy decisions accountable, might occur. If this same spirit of openness were to become the standard in higher education, the next thing is that one would see people underlying the profit motive of education by developing Bayesian analysis tools and open-sourcing them. Or there might even be some mad few who offer their expertise online to allow anyone in the world with access to internet the opportunity for education and inspiration. I will never pay it (the privileges I enjoy) forward (to those who follow). As we say in my part of the world, “Yeah, right”. http://www.tui.co.nz/competitions-and-events/yeah-right-gallery

- *Llewelyn Richards-Ward* says:

September 15, 2016 at 11:55 pm

“undermining the profit motive”…

3. *Rahul* says:

September 16, 2016 at 12:02 am

I think the issue is not so much the charges themselves but unreasonable charges. e.g. FOIA requests get charged too but $300/hr seems ridiculous.

Is the Flourishing Families project staff working on this request getting an annual paycheck of $600k?

- *elin* says:

September 18, 2016 at 10:51 am

Agree, there are definitely costs involved in doing the work (I used to be the person who would get pulled off of other work to deal with that stuff and also people with questions about our ICPSR data) and I can say there is no question that it is an imposition both as distraction and in time. But this fee is incredibly high. They should be doing something more like the “federal rate” for consulting.

  - *Rahul* says:

September 19, 2016 at 1:44 pm

What is the federal rate for consulting? Just curious.

4. *Shravan* says:
   September 16, 2016 at 3:37 am

   Andrew, how about putting up the data and code alongside the link to your papers on your home page? That is what I do. I have also started keeping data and code on github, i.e., publicly.

   What these people are doing (charging money for the preprocessing) is unethical and is probably under false pretences anyway. How did they analyze their data without doing any preprocessing? 80% of data analysis is data preprocessing. Are they charging for stuff they did already before they did their analysis in order to publish their paper? That's what I want to know. If I were receiving such a bill, I would want to find that out first.

   - *Andrew* says:
     September 16, 2016 at 8:18 am

     Shravan:

     Yes, posting data and code would be a good idea. For one thing, then it would be easier for *me* to find this information when needed.

   - *Keith O'Rourke* says:
     September 16, 2016 at 8:38 am

     It is likely that the data was not properly archived and it will cost them to put it together and I would not be surprised if it cost more that $450.00.

     A group I used to work with spent over a month struggling to extract and put together control group data from three studies the group had done and published just a few years earlier. I was told that they had finally succeeded and the data was put on a diskette for me (OK this was around 20 years ago) and given to the senior academic to review first who took it home and apparently lost it somewhere in their basement. They also informed me that unfortunately no one had thought to make a backup copy of the file. I never got the data. Now to give a sense of the academic success of this group, the senior academic has 200++ publications with 50,000++ citations.

     - *Shravan* says:
       September 16, 2016 at 8:49 am

       Interesting. Incidentally, the majority of researchers in psych* are unable or unwilling (I don't know which for sure) to release data. Has someone tried to get Amy Cuddy's data? It would be fun to reanalyze the data from her key paper that made her so famous.

       - *Andrew* says:
         September 16, 2016 at 11:24 am

         Shravan:

         I don't know that even Amy Cuddy has direct access to Amy Cuddy's

data. Given all the simple calculation errors in these papers, I'm guessing their record keeping is a mess.

- *Greg Francis* says:
  September 16, 2016 at 11:43 am

  At least some of Cuddy's data is at

  https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FMEGS6

5. *BoSelecta* says:
   September 16, 2016 at 3:50 pm

   I've been following a few psych forums where active scientists post. You have no idea how offensive they find the idea that someone might want to have a look at their data and how forgiving they are to all those who want to charge others for dragging and dropping a few files.
   What more fascinating is that they argue, that perhaps the data was not tidy and sharing it required cleaning it up. If you can publish a paper based on data which hasn't been cleaned…

   - *jrkrideau* says:
     September 17, 2016 at 7:36 am

     If you can publish a paper based on data which hasn't been cleaned…

     Well do so carefully?

     Actually I think I understand the point. I have had datasets cleaned and usable but which, somewhat stupidly on my part, required researcher knowledge to use effectively.

     There can be a fair amount of effort in documenting some of the rationales for data cleaning, information on where specific numbers come from, that other people can use the data with some confidence.

     I inherited a data set and a mess of SAS programs that, when I handed them over to a new custodian, it took me 3 weeks working full-time to document the files and assumptions made in the analysis.

     I have since learned to, at least, put a lot of verbose comments —for example data from StatCan survey XXX Table 2 –in my R-code since the programmer's rule applies to data analysis as well—the original programmer is a new user after 6 months away from the code.

     - *elin* says:
       September 18, 2016 at 10:56 am

       Yup … actually one of the issues based on my experience is that if you include all of the cleaning code everytime, and then sometimes you are adding recodes or creating scales and other times not, depending on what

you are doing, it's actually quite messy. I really prefer getting a common, clean version that you load when doing the analysis. However, when asked to share data it is unclear what that means and how much cleaning you should include.

6. *Bo Chen* says:
   September 16, 2016 at 5:21 pm

   Data is an important resource in my field Agricultural Economics and use of many datasets are restricted under agreements with data providers such as companies and government agencies. It is simply illegal to share data freely.

   - *Rahul* says:
     September 17, 2016 at 9:12 am

     Which raises a good point: Should open data be a necessary pre-condition to publishing in Journals or not?

     - *Martha (Smith)* says:
       September 17, 2016 at 11:59 am

       @Rahul

       A tough question. Open data is the ideal, but in some cases such as those Bo Chen mentions, it is not possible (sometimes for poor reasons such as dubious "trade secrets" and sometimes for good reasons such as protecting subjects' privacy when sensitive issues are involved). Definitely open data ought to be provided unless data non-disclosure agreements were required for the authors to do the analysis. But there needs to be some flagging of papers whose data is restricted by such agreements, along with caveats that the authors have no information on the quality of the data collection, so results are contingent on such quality. And there needs to be transparency on what data collection methods were (and on the lack of such information when restricted, and caveats related to such lack of transparency).

       - *Rahul* says:
         September 17, 2016 at 12:55 pm

         Would the downside be big if work such as Bo Chen's just remained secret? I mean there's tons of unpublished, trade secrets. Let this remain so too?

         What irks me is that in the absence of a strict requirement for open data as a precondition to publishing isn't it hypocritical & ad hoc to call out individual authors for not sharing data?

         - *Andrew* says:
           September 17, 2016 at 8:39 pm

           Rahul:

People can do what they want to do. But if someone wants to charge $450 for their data, we can sure as hell make fun of them!

- *Rahul* says:
  September 17, 2016 at 10:29 pm

  

  “Progress by clear rules” versus “progress by mocking”?

  - *Andrew* says:
    September 17, 2016 at 10:33 pm

    

    Why not both?

  - *Rahul* says:
    September 17, 2016 at 10:58 pm

    

    Both is good. Why don’t senior academics take a stand:

    Insist on open data policies at all journals you are an editor at. Step down if they won’t play ball. Don’t let Journals pay lip service to open data by having a policy they don’t enforce. Decline to review papers for Journals that don’t insist on open data. Refuse to submit papers to Journals that won’t insist on open data.

    In general, I see a lot of mocking & complaining among academics on this issue but no hard action.

    - *Keith O'Rourke* says:
      September 18, 2016 at 9:44 am

      

      Agree, then we might avoid these sort of messes http://www.medscape.com/viewarticle/832483

      (I have no first hand knowledge who is right or wrong here but I did work with the authors of the paper in the past and my prior puts more probability on them being wrong than right here.)

- *elin* says:
  September 18, 2016 at 11:09 am

  

  Lots of individual level data from surveys or other records is also restricted because of the ease of back-indentifying individuals. That’s why the NELS88 data which includes individual transcript is not allowed to be used on a computer that is connected to the internet, just as one example. ANd of course all detailed census data that is less than 75 years old can only used at an authorized repository with serious justification. If I know that someone had an HIV diagnosis in town x in 2009, that person may be easily identifiable.

  I also have been in a situation where I submitted a data set to ICPSR and later

learned about another data set that could easily have been matched up with the records for about 10% of the sample. At ICPSR though, users have to agree that they will not attempt to back-identify or match.

There can be major ethical issues and acting like they are just nothing is not particularly useful.

- *Martha (Smith)* says:
  September 18, 2016 at 4:24 pm

  +1 This is why we need exceptions to data-sharing requirements — but need to delineate clearly what the good reasons for being an exception are. And we need to promote re-analyses by others of studies where there are good reasons why data cannot be made public.

  - *Keith O'Rourke* says:
    September 19, 2016 at 7:53 am

    > promote re-analyses by others of studies where there are good reasons why data cannot be made public.
    That would be a fix that may often work, a second independent re-analysis by qualified others.

    Here is a case study of that not working very well http://www.hc-sc.gc.ca/dhp-mps/medeff/advise-consult/eap-gce_trasylol/final_rep-rap-eng.php (search for satisfactorily explained and see my earlier link).

7. *Martha (Smith)* says:
   September 17, 2016 at 6:08 pm

   Rahul said: “What irks me is that in the absence of a strict requirement for open data as a precondition to publishing isn’t it hypocritical & ad hoc to call out individual authors for not sharing data?”

   I guess that depends on what you mean by “strict”. For example, if there is a stated policy that data are to be shared unless specified conditions for non-sharing apply and are documented, then it it not hypocritical to call out someone who does not share data when those conditions are not met or not documented.

8. *Chip Lynch* says:
   September 19, 2016 at 12:47 pm

   Just got back from International Data Week, in Denver, Colorado where the theme was “From Big Data to Open Data”. So a good portion of the sessions were related to Open Data and sharing and these sorts of issues. Too much to cover in a comment, but for anyone interested there were lots of great talks and the content is mostly online, with some videotaped sessions apparently being posted later. The main site is here: http://www.internationaldataweek.org/ (Sorry if this sounds like a shill… I’m not involved, just a fan; it was one of the better conferences I’ve ever attended).

   - *Rahul* says:

September 19, 2016 at 1:43 pm

Does Open Data have private / corporate interest in the way that Big Data did?

Just a random thought. Trying to see if there will be ways in which it is in private enterprise's interest to go Open Data. If not, then this may be a handicap Open Data has that Big Data didn't?