Feb 16, 2012

Paying for [no] sex

My friends Berk Özler and Sarah Baird (and coauthors) just published a very nice paper in *The Lancet*. They study adolescent girls in malawi, where the randomized intervention is cash transfers, either unconditional or condition on school attendance. They find that, compared to a control group which received nothing, the cash transfer group (whether conditional or not) had significantly lower rates of both HIV and HSV-2 (i.e. herpes). See also a nice write-up in *The Economist*.

In the meantime, I and many coauthors have a new paper in *BMJ Open*, an open access spin-off of the british medical journal. We study both male and female young adults in tanzania. The intervention is a conditional cash transfer, where participants only receive the money if they test negative on a suite of STIs (chlamydia, gonorrhea, trichomonas, etc). Anyone who tested positive received free treatment and was eligible for the reward at the next testing four months later, which is one reason (along with low incidence rates) that we did not concentrate on HIV as an outcome. We find lower incidence of the suite of STIs in the treatment group as compared to a control, but only at 12 months (not at 4 or 8 months) and only for the high-reward arm ($20 per round vs. $10 in the low-reward arm). The fact that we only see an effect at 12 months is not particularly surprising or problematic, since it may have taken some time for participants to trust the fidelity of the intervention. However, it was quite surprising to me that we only saw an effect with high rewards, since my guess would have been that any non-trivial amount would be both incentivizing and (probably more importantly) salience-enhancing.

Note on publishing: we had submitted the paper to the BMJ itself, who gave it guarded reviews and asked for a revision. We responded and were hopeful that it would be accepted, but sadly the further reviews were not sufficiently positive and the editor rejected it. Fortunately, they immediately solicited it for BMJ Open, where it was accepted on the basis of the existing reviews. What makes this all somewhat unusually interesting is that everything is public: the reviews (including names) and our response. See the full history here. One amusing critique that particularly caught my eye is in the second-round review:

> Calculation of P-values between randomized groups at baseline is illogical – the P-values indicate the probability that differences have occurred by chance – as all differences are created by randomization, they must have occurred by chance, so why calculate a probability? What is important is the magnitude of the differences, not the P-values. Please remove the P-values from Table 1 and the baseline data section of the results.

Most readers here probably won't care, but this was jarring to read since the practice is quite standard. On the one hand, he's right that it's a bit silly to calculate p-values when we know full well that assignment to treatment was perfectly random. And some of you will remember that I am the first to implore folks to focus on magnitudes rather than p-values.

But even I think this reviewer may be asking a bit much. Technically he's wrong, despite being a professor of biostatistics: the p-value tells us the probability that a distribution of numbers at least this far apart would be observed via random draws conditional on the null hypothesis that the treatment and control groups have the same mean. It is only a property of the numbers, not of how they were actually derived in the real world, so it is perfectly meaningful even here. More saliently, but on a related note, the p-values are still a useful shorthand (along with the magnitudes, of course!) for whether the observed differences are 'unnaturally' large, and they are easily and helpfully interpretable in that vein for many people.

- Digg this post
- Recommend on Facebook
- Tweet about it
- Print for later

Category: Economics, Research

8 Responses

1. *Gary Collins* says:
   February 16, 2012 at 5:37 am

   Just because something is 'standard practice' for some journals (e.g. NEJM) doesn't make it the right thing to do. Suggest you read the CONSORT Explanation & Elaboration paper and focus on item 15 the section on reporting baseline data

   http://www.consort-statement.org/consort-statement/13-19—results/item15_baseline-data/

   Look at some trials reported in JAMA and Annals of Internal Medicine, you tend to find no p-values in Table 1.

   Reply

2. *Julian Jamison* says:
   February 16, 2012 at 8:43 am

   Thanks for the comment Gary, and for the linked content which I hadn't seen, but the fact that the reviewer quoted Consort guidelines doesn't make his argument any more convincing. [Btw, I wasn't referring to NEJM, but rather to essentially all journals in economics and many other social sciences, where I come from.] I agree that "standard practice" doesn't make something right, just like being in Consort (or JAMA) doesn't make something right.

   I give a specific statistical explanation and rationale in my post above; do you disagree with it on substantive grounds, and if so why? Let me give yet another rationale here: sometimes even after randomization, there are several additional steps before enrollment in the treatment group. Even if in theory these should go smoothly (and we try to keep track of anyone lost along the way, so they still count as "treatment"), things can always go wrong. This is especially true in rural tanzania or malawi, as opposed to major hospitals in the US or UK. So it is in fact quite possible that inferential rather than descriptive tests are necessary at baseline to assess whether that could have occurred.

   Reply

3. *Gary Collins* says:
   February 16, 2012 at 9:42 am

   CONSORT is the product of years of intellectual effort to produce consenus amongst methodologists, clinical trialists, journal editors and clinicians (see http://www.consort-statement.org/about-consort/history/ for history) to improve the quality of reporting of randomized trials etc…it wasn't derived on the back of an envelope. Hence it's wide endorsement in over 600 journals, International Committee of Medical Journal Editors (ICMJE), World Association of Medical Editors (WAME) and Committee on Publication Ethics (COPE).

   There is no rationale on why you should ever test for baseline imbalance – just put in place a good randomization procedure – this issue has been discussed on numerous occasions over the years in the medical literature and the overwhelming opinion is not to test for baseline imbalance – I suspect more so than in the economics and social sciences. Have a look at

   Senn S. Testing for baseline balance in clinical trials, Stat Med, 1994; 13: 1715-1726.

   Any strong prognostic factors should be dealt with accordingly in the randomization procedure to ensure any imbalances are minimal and key prognostic factors (or factors included in the randomization procedure) should be used in the analysis.

   Spurious (by chance) differences in baseline characteristics (identified by testing of baseline differences) in studies where the intervention has not been shown to offer any improvement over the standard approach are then often used to justify why the treatment was not shown to be beneficial, which is totally false or to dictate how the analysis is to be carried out, which should be specified upfront before the trial closes and definitely without looking at the data!

   Reply

4. *David McKenzie* says:
   February 16, 2012 at 1:47 pm

   Nice discussion Julian. I'm with Gary and your referee here – my paper with Miriam Bruhn "In Pursuit of Balance" has a section discussing concerns with reporting p-values in a Table 1, including the potential for abuse. The working paper version (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1293165&download=yes) has more, including this paragraph:

   "Overall, in most randomized settings, we therefore recommend reporting the point estimates of the differences across groups, but not reporting the p-values on these differences. If these differences are in variables thought to influence the outcome of interest, one should control for them, regardless of whether or not the difference is statistically significant. Note, however, that there are two exceptions where carrying out a test of statistical significance is meaningful. First, statistically testing the difference between treatment and control groups at baseline can be relevant if there was possible interference in the randomization. This may be relevant when random assignment is carried out in the field by survey enumerators, but should not be a concern when the researcher does the randomization by computer. Second, a related common use for these significant tests is seen in Ashraf et al. (2006a), who are only able to survey 1777 of the 4000 microfinance clients allocated to treatment and control. They test whether there are differences between the treatment and control groups amongst those surveyed."

i.e. when there is a risk of interference with randomization or attrition from the subject pool, this seems like a useful test, but otherwise not so much.
David

Reply

5. *Julian Jamison* says:
February 16, 2012 at 3:32 pm

First, apologies for my initial reply, which was I think a little testy (too early in the morning for me, and too rushed). Second, thank you both for contributing. I love to think about this kind of thing (math nerd), and don't know much about it (obvious?), so I'm learning a lot. The references are very helpful — I had seen David's paper on how to randomize, which everyone in our field should pay attention to, but hadn't noticed the section on testing for balance at baseline.

All that being said, I'm still deeply unconvinced. Just because Consort says it, no matter how much thought went into that, doesn't make it right. The main argument seems to be that since there is potential for misuse, we should not include the results of certain statistical tests… but there is potential for misuse with everything in statistics, so that's not going to get us very far. The Senn paper (which I read in full) and for the most part the relevant section of David's paper are primarily concerned with using such baseline tests to decide which covariates should be included in the final analysis. E.g. Senn says: "The general message of this paper is that covariates are not to be fitted according to whether or not they are 'unbalanced' but according to whether or not they are important." I agree fully!!

But that implies that you shouldn't misuse baseline tests of balance, not that you shouldn't use them. Senn also says: "I am prepared to concede a minimal role for a test of baseline balance as a test of randomization." Well, yes. That's the whole point. David explicitly states that (especially in the field) randomization is not always assured. I would go a step further: even if randomization is done on a computer, that doesn't mean it matches with what happens in the field.

For instance, I'm working on a study in namibia where we randomize at the community level. We collected community-level data on lots of variables before deciding which communities were eligible to be in the sample; then we randomized communities; then we collected the individual-level baseline data. Obviously it would have been better if the latter two steps had been reversed, but our implementing partners needed to know asap where they would be working (in fact they wanted to know months before we could tell them), and these areas are remote so we couldn't simply pop out and survey everyone in a week. Given that the enumerators might therefore have gleaned which communities were in the treatment group, it's possible that they (unconsciously?) interacted differently with them, leading to imbalance at baseline. I frankly think this is extremely unlikely, but I'd like to be able to test for it and convince the reader. My point is that you can't be sure of anything, so having more information, properly interpreted, can only help.

Finally (I promise), I still think the p-value is simply a nice shorthand to help the reader get a sense for what the data look like. For instance, even if you specify important covariates in advance, "controlling" for them by putting them into a *linear* regression (recall that all models are mis-specified!) doesn't get rid of their effect. [David: I liked the relevant simulations in your paper, but you can't prove that it will always go away.] Better if they're balanced at baseline, and the test of balance – in addition to the magnitudes – helps convince the reader if this is the case.

Reply

6. *Larkin Callaghan* says:
February 17, 2012 at 12:50 am

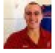
Looking forward to reading the whole article (just checked the abstract) – my first question was does it deal with/how does it deal with STIs transmitted via sexual assault or abuse?

Reply

7. *Julian Jamison* says:
February 17, 2012 at 2:51 pm

Larkin – you ask a good question and (implicitly) raise a fair criticism. Everyone (treatment and control) was given sex ed and access to sexual health services. However, there was nothing specific done to control for sexual abuse (or indeed being in a relationship where one partner was not in a position to influence e.g. condom usage). In that sense it "unfairly" punished those respondents. Of course, part of the goal (as in the malawi paper I cited) was to see if access to money would help people leave such relationships, or change the power dynamic within them. We hope that was the case, although of course we can't prove it either way.

Reply

8.  *Jason Kerwin* says:
February 17, 2012 at 8:13 pm

Julian's initial statement about the definition of a p-value was exactly right, and I think contemplating it helps clarify why it's good idea to include them:

"The p-value tells us the probability that a distribution of numbers at least this far apart would be observed via random draws conditional on the null hypothesis that the treatment and control groups have the same mean."

That is, let's assume we actually did have perfectly random assignment. Random assignment does not guarantee *even* assignment. By summarizing the distance between the two group means, p-values clarify the extent to which we got an unlucky draw – one that is very close to the extremes of the probability distribution.

Ed Leamer has a great example of this in "Let's Take the Con out of Econometrics" (http://www.international.ucla.edu/media/files/Leamer_article.pdf). He imagines a farmer testing the effectiveness of a fertilizer. The farmer randomly assigns each of his plots to either get fertilizer or not. But there's some chance that all of the fertilized plots will end up under trees. The randomization doesn't prevent that from being a problem. It only guarantees balance (on all covariates, no less!) in infinite samples.

The point about not using Table 1 p-values as a guideline for eventual regression controls is a good one. Maybe it's best to hide them if people would be inclined to use them that way. But we should really control for *all* baseline characteristics, and doing so won't cure all our ills if there's a complex relationship between a baseline characteristic and our outcome.

Reply

Leave a Reply

Name (required)

Mail (required)

Website

Submit Comment

□ Notify me of followup comments via e-mail. You can also subscribe without commenting.

Julian Jamison

I'm an economist, researcher, traveler, runner, and astronaut-in-waiting. I enjoy pondering human behavior, including both what we do and what we ought to do - either to maximize our well-being or in pursuit of some other goal.