CONVERSATIONS ON DATA SCIENCE

ROGER D. PENG
HILARY PARKER

Conversations On Data Science

Roger D. Peng and Hilary Parker

This book is for sale at
http://leanpub.com/conversationsondatascience

This version was published on 2016-08-06

This is a Leanpub book. Leanpub empowers authors and publishers with the Lean Publishing process. Lean Publishing is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

Contents

Not So Standard Deviations: The Podcast

Thanks for purchasing this book!

The content in this book is based on episodes of *Not So Standard Deviations*, a podcast that we started in 2015 and continue to publish about every two weeks. On this podcast, we talk about the craft of data science and discuss common issues and problems in analyzing data. We also compare how data science is approached in both academia and industry contexts and discuss the latest industry trends.

You can listen to recent episodes on our SoundCloud page or you can subscribe to it in iTunes or your favorite podcasting app. We are also available on Stitcher and through the Google Play Music store.

If you want to support the podcast directly, you can become a patron at our Patreon page. For $2 per episode, you can get a really cool *Not So Standard Deviations* hex sticker!

If you have any feedback for us about the podcast or the book, or if you have any questions you'd like us to discuss, you can email us at `nssdeviations@gmail.com` or tweet us at @NSSDevations.

Thanks again for purchasing this book! We hope that you enjoy it.

Roger Peng and Hilary Parker

Analyses that Seem Easy

Power and Sample Size Calculations

Roger

I want to talk about analyses that seem easy, but end up being hard. This comes up often for me, but one of the things I want to talk about was the power and sample size calculation, which is the bread and butter of the biostatistician.

Hilary

And also, web company–the ones that are doing things right.

Roger

I'm interested to hear what you have to say about this. In my job, the way it works often is collaborators come to me, and they're designing a study. It might be a clinical trial, it might be an observational study, and they just want the sample size calculation (i.e. how many people they need to enroll in the study). Good collaborators give you a couple weeks' notice, and bad collaborators give you a couple of hours' notice. Often there's some grant deadline that's pending and they need to write this for a grant.

Very often, the presentation of the problem comes off as, "I know this is really easy. Can you just do this really quickly and just give me a number?" Often, the sample size is just determined by the budget so that's fixed, and then they want to know what the power is going to be or the estimated effect size is going to be. So there's three things, "What's the power? What's the effect size that we can detect? And then what's the sample size?" Actually, I least often calculate the sample size. More often, I calculate the effect size.

Hilary

That's the right way to go.

Roger

As Karl Broman once famously said, "The sample size is equal to the cost per sample divided into the budget."

Hilary

Yes, when somebody comes to you for a sample size calculation you say, "Well, how much money do you have?"

Roger

On one end, you might say, "Sample size is a function of the budget." But another way to say it is, "Sample size can help you determine whether the question you're talking about is reasonable or totally infeasible."

If the sample size is 100, then it may be feasible, but if it's 1,000, then it's not, at least in my context.

Hilary

That’s my biggest qualm with…I know from having worked in stats, this happens all the time where someone comes to you too late, they’ve already designed their experiment and haven’t decided what they are going to do, and then they ask for a sample size calculation at the end, and you have to say, “Great, you can run this experiment for 87 years. You’ll see the effect size that you’re interested in.” And they’re coming so late that you feel like a wet blanket.

Roger

It’s always horrible to deliver that kind of news because the question that you focused on for weeks is totally infeasible.

I feel like the less-prepared collaborators come to you and say, “Everything’s done. We just need you to stamp this number and say that you approve.” The better collaborators, rather than saying, “Hey, I need a number,” or, “I need a power calculation,” or, “I need an effect size,” they say, “We need to have a conversation about the science and we need to know what’s feasible and what’s not.” At a high level, you need to know in terms of order-of-magnitude-type of questions, in terms of feasible or not feasible. Then when you know what’s feasible, you can figure out, “How can we optimize so that we ask a question or we look at the effect that we can get the most juice of in terms of the budget and how much we can afford to do.” Do you have the same experience?

Hilary

I think one thing that’s sort of interesting is that, because experiments are such an ideal, I feel like we lose sight of the fact that they are *the* gold standard, (and the best thing you can do if you want to find causation) and that there are alternatives out there. So if something's infeasible, it doesn't mean you can't do *any* science whatsoever. You can still do analysis. There are statistical methods to deal with imperfect setups. I've found that conversation is hard to train people to have. They'll either come to you and say, "We want this number, and if this number doesn't work, then let's throw the whole thing out."

I feel like the conversation that I wish I had more that I don't is, in many companies, you can train engineers and product managers to understand that experiments are important and sample size is an important aspect of doing the right type of experiment. But if the sample size calculation comes out and it's a very low-traffic page or something, and you get to this situation where you can't get something perfect, there are still so many options in-between doing nothing and doing an A/B test or an experiment.

I find it hard to get people to have that conversation at the right moment in time. I think it's sort of the flip side of teaching people to do the perfect thing is that there's sort of less area for gray or in-between, and that's somewhere where, I think, in academic science or statistics is maybe more of an understanding of that.

Roger

I have to think back on that. One thing I've found—and maybe academia is a little unusual in this way because it's supposed to be doing things that are kind of "different"—but I've never had a totally routine power calculation. I feel like every time someone's needed one or I work with someone, there's always been a couple of things that make this problem unique, you know?

Hilary

I think that is this difference I was talking about. In a tech company, if you're changing something on a website and then you decide to change something else a week later, you can follow the exact same procedure. In a tech company, it's easier to make things standardized, which is why I think experiments have sort of flourished in that environment.

Roger

It's easier to control things.

Hilary

It's so much easier. And then traffic is "cheap", or it's easier to get sample size. Samples are very cheap to acquire, usually. It's one of the things where it's, "Why not just throw the gold standard at it. It's relatively straightforward and cheap to get. Experiments are cheap." Whereas in medicine or in academia, in most academic applications, samples are much more expensive to acquire and so there's all these methods to take account for that. I kept running into this problem where if you can't do the perfect thing, people don't understand that there are statistical methods for the imperfect approach.

There's a lot of work being done on causal inference, for example, people at Facebook are looking at that problem right now. That's something where that is almost a niche thing in tech companies and the standard is these perfectly-implemented experiments. It's sort of the flip problem from your research in environmental health—you can't run an experiment on temperatures.

Roger

No, it's definitely not the norm.

Hilary

You run into a different set of problems when you try to apply experiments at scale. It's been interesting to see the flip side, and it's all driven by this cost of samples.

Roger

Now that I think about what you're saying, it's seems a little weird that in academia, I feel like all the time, people are scrapping whatever data they can find because running an experiment on people is just so expensive. So they just kind of gather whatever data they can get, and then you run into all these problems in terms of the analysis and drawing causal conclusions, whereas you guys have all the data coming in from wherever. You're more conditioned to do the controlled experiment in a non-academic environment.

Hilary

That is true. So conditioned to do that that there's not necessarily an understanding that this is this gold standard, perfect way of analyzing causal inference and that there are other options available. That's been super interesting for me.

Roger

It's just a function of money, obviously, but I would have thought that the thinking would have been the other way around.

Hilary

It runs into its own set of problems because there's sort of an all-or-nothing attitude. You either just look at traffic and understand things, or you can do an experiment and understand whether or not it was causal. But then doing something in-between, like propensity score matching or something like that, isn't an option. I feel like when you look at web experiment or web analysis, it's one pole or the other and there's not as much in-between except from these large tech companies.

Roger

I actually kind of have a nerve-wracking power calculation.

Hilary

Do tell.

Roger

I was on one of these multi-center clinical trials and they had an interim analysis. Normally in a clinical trial, you're not allowed to look at the data, or at least the outcome data, until it's over because otherwise, that might bias your implementation of the study. This is a five-year study and in year three or so, there's an interim analysis planned where we would look at the data half way through and determine whether or not the experiment was having an effect. Usually, the idea is that if it's having this dramatic effect, then you would stop the study because then the control group is not getting the treatment, and it's not really ethical.

One of the issues that also can come up is if there's no effect, then you have to determine whether or not continuing the study will allow you to observe an effect. It's conditional power for futility. If you think there's going to be no effect, will gathering more data allow you to see the effect? If gathering the other half of the sample won't allow you to see it, then you can draw this kind of idea that it's futile. Anyways, I feel like I've never had so much riding on one calculation, you know?

Hilary

Yeah.

Roger

It's not like they make the decision based on that one thing that I say, because there's other data that they look at, too. But it was kind of interesting to go through that process for the first time.

Hilary

Were you dusting off the derivation of the different...

Roger

I was like…yeah, “I really better get this calculation right this time, unlike all those other times.”

Hilary

I find power calculations deceptively hard because no one ever talks about what it’s powered to, like you’re detecting an X% change in the effect. There’s that concept of the difference that you’re powered to observe 80% of the time. I find that part of the reason why it’s deceptively hard because the moment you start actually digging into it, if you’re discussing it with someone, there’s all these things that they weren’t expecting to even come up, like judgment calls you made about the false positive and false negative rate.

Roger

In my experience, it’s very difficult for people to say in most cases what a practical effect size is.

Hilary

That’s a very common question that comes up at Etsy, “What effect size should we be seeing? And everyone comes to the statistician thinking, “This is a statistics question.” It’s not.

Another one that comes up a lot is what correlation means “something”, and I’m like, “Trust me, if you’re in physics, correlation of 0.8 would be terrible, but if you’re in social science, that would be amazing.”

Roger

I've found that people will usually have some upper or lower bound. If I just say something ridiculous like, "What about a 90% change," they'll be like, "Oh, no, no, that's too big." I usually try to bracket the interval somewhat, and then if I can get it within the reasonable amount, I'll just take some middle value, and they'll be okay. I know it's hard because for a lot of health-type questions or biological questions, you just don't know. The human body is just too complicated. So you don't really know what a meaningful change is going to be. And so you just have to take a stab at it. In that case, as a statistician, I'll just make up the number. That's why the effect size is always the easiest thing to calculate because no one ever knows what it is.

Hilary

Especially if someone's saying, "We have X dollars to spend." Realistically, we could never have more than 100 people in this study. Then that makes it easy. I will do that here, too. I will say, "Okay, there are certain constraints on the type of experiment we run. I'm going to present you with the maximum detectable effect size that you're going to see. So manage your expectations accordingly."

A/B Testing

Roger

The other thing that I thought sounded like it was straightforward to do, but probably actually hard, is A/B testing or what we call clinical trials. I think people understand that clinical trials are hard because you've got to enroll people or whatnot. But A/B testing in a tech company, it seems like it would be easy. You've got data coming in all over the place. You're testing two things. What could be so hard about that?

Hilary

I think there's a sort of attitude that it's very easy. A lot of data scientists will have this attitude that it's not that interesting of a problem. I think it's super interesting. I also come from biostat so that makes sense. But it seems like it's very simple, "Oh, just count on one side versus the other and then you're done." I think the complexity comes from the data-generating process.

There's all sorts of judgment calls you make, like what constitutes traffic, a visit. If you have visits, and if you have multiple people visiting multiple times within one experiment, if they're always in the same variant, then you have correlation within the visits, right? There's things like that that. The web traffic gets so oversimplified before you even see it that the A/B tests always get results. You'll be able to do a simple calculation like a proportions test, but then you won't be able to explain strange patterns in the results. When you start digging, it becomes a horror show.

Roger

When you were talking about that just now, I already started to sweat.

Hilary

The number of times I've said this personally, that the i.i.d. (independent, identically distributed data) assumption for proportions tests is almost always violated, either the independence, or the identically distributed part. It's almost always violated when you're doing A/B testing. Then it's a question of scaling—how much the broken assumptions are affecting it. T-tests can be robust, but it is not as simple as it sounds.

You have to choose when someone shows up what variant to show it. So you have to do that via a pseudo-random process. Even that, I'm sure you're kind of immediately thinking, "That is not perfect," right? So even the way that you're bucketing people isn't perfect, and then those get rolled up into visits, which is an imperfect kind of arbitrary thing. The way visits are defined by Google, that I think a lot of companies use, is it has to be some action on the site. So some event is generated, and the time between the events is less than 30 minutes. So once two events happen that are more than 30 minutes apart, those are 2 separate visits. So it could be someone who got up and had lunch and sat back down and was on the website. It's a very imperfect.

Roger

And 30 minutes, you just made that up, right?

Hilary

Google made it up at some point

Roger

And that's like a 0.05. It's never going to change, right?

Hilary

Exactly. Google is the Fisher because they're the ones that just made this arbitrary decision that has affected the field tremendously. Because the reason why companies will use that definition is because most companies will have started with some version of Google Analytics on their site, and then they're like, "Well, we want it to be apples to apples with this old system so let's just continue to use the same definition, and that makes thing simpler."

Roger

It seems like every company probably starts with Google Analytics, right?

Hilary

Google makes it very easy so a lot of people use it. I had Google Analytics on my academic site, actually, might still. I haven't looked at it in like a really long time. But they make it so easy–just paste in code, and you get it.

Roger

I know. It's a little too easy. And also, a lot of service providers now, it's all integrated. So you just type in your ID and it just goes.

Hilary

When I'm talking about this problem, I always say, "Have we given people just enough rope to hang themselves with." If you are giving people these very simple, digestible statistical testing, and they don't have any idea or they've never been educated about all of these caveats that are going to make this testing so simple, that makes it really hard to untangle when things start to look weird, which invariably will happen when you have these weird, complex data structures that might end up breaking.

I think it's a really cool field. I'm definitely focusing more and more on experiments, and I think it's a fun place to be a statistician because you're thinking about the same problem in a different way with a different set of constraints. But the most frustrating part is that there's this attitude, "Oh, it's just A/B testing. That's easy."

Roger

So one thing...I will just reveal my total ignorance of this area, but I'm going to say it anyway. One thing that you said that kind of like a light bulb went off is when you said when the people come into the site and you have to assign them to a group. And let's say for the sake of argument, there's two groups, right? And I always just figure, "Well, there's some random number generation process going on in the background and you flip a coin and you get heads or tails." But you said, I guess, you hash something that's unique to their visit and then...

Hilary

It becomes deterministic for the visit.

Roger

It's always deterministic, ultimately, but I guess it occurred to me that you have all these instances and because you have so many people coming at the same time, right? I guess it would be hard to synchronize a random number generator across all these instances, right? I was thinking, "Why do you hash something? Why didn't you just have some coin-flipping in the background?"

Hilary

Yeah, that's the issue.

Roger

I'm sure it's more efficient, too, to take some piece of data from the user and turn it into their assignment.

Hilary

Yeah. I mean, the thing that I have learned from working at a tech company is that whatever a statistician would do, some data engineer has done something much more efficient but much more complicated-sounding to simulate what the statistician says.

Roger

I imagine, in many cases, that's a function just like the realities of the load and whatever. I think what Christopher Volinsky said that when he came to AT&T, I think Daryl Pregibon told him basically, "Everything that you've learned, forget about it." It's a version of that.

About the Authors

Roger D. Peng is a Professor of Biostatistics at the Johns Hopkins Bloomberg School of Public Health. He is also a co-founder of the Johns Hopkins Data Science Specialization, the Simply Statistics blog where he writes about statistics for the general public, and the *Not So Standard Deviations* podcast. He is the recipient of the 2016 Mortimer Spiegelman Award from the American Public Health Association, which honors a statistician who has made outstanding contributions to health statistics. Roger can be found on Twitter and GitHub at @rdpeng.

Hilary Parker is a Data Scientist at Stitch Fix and co-founder of the *Not So Standard Deviations* podcast. She focuses on R, experimentation, and rigorous analysis development methods such as reproducibility. Formerly a Senior Data Analyst at Etsy, she received a PhD in Biostatistics from the Johns Hopkins Bloomberg School of Public Health. Hilary can be found on Twitter at @hspter.