The Chance Manifesto

Peter G. Doyle

DRAFT Version 0.3
2 June 2004

Aren't you the lucky one?

Once upon a time, back when I was in grad school, I ran across a misshapen die. (Here 'a die' means 'half of a pair of dice'.) Something had obviously gone wrong in the manufacturing process, and instead of coming out nice and cubical, this die was a rectangular parallelopiped with side-lengths (a, b, c), $a > b > c$.

Now, as you probably know, the sides marked 1 and 6 on a die are invariably opposite one another, and in this case the 1 and 6 were marked on the two $a \times b$ faces. So this die had a tendency to come up 1 or 6, in preference to $2, 3, 4, 5$.

Or so it seemed to me, after rolling it a few times. But perhaps I was being too hasty. Maybe this die, though it appeared as if it should favor 1 and 6, was really close enough to cubical so that its behavior would be almost identical to that of a standard die. Maybe the preference for 1 and 6 I claimed to be seeing was just a statistical fluctuation.

I decided to carry out an experiment. I would roll the die a bunch of times, count how many times each number appeared, and look at the resulting distribution.

So I rolled the die 100 times, and kept track of the results. As I had expected, sides 1 and 6 were the two most frequent outcomes. I concluded that the die did indeed have a preference for coming up 1 or 6.

QUESTION: Do you think this die really did have a preference for 1 and 6?

Well, let's think: For a standard cubical die, the probability that 1 and 6 will be the two most frequent outcomes in a given number of rolls is roughly

$1/\binom{6}{2} = 1/15$. (Only 'roughly', because sometimes there will be ties between the frequencies.) So having 1 and 6 come up more frequently than $2, 3, 4, 5$ could have happened just by chance. A probability of 1 in 15 is pretty substantial. Really, we'd want to see just how much more frequent 1 and 6 were than the other outcomes, before we buy into this idea that the die was 'loaded'. Maybe we could get hold of the raw data, and do our own statistical tests? On the other hand, there was presumably good reason to think that the die was loaded. How good the reason was, we don't know, because I haven't revealed just how misshapen the die was, presumably for good reason.

Other aspects of this 'experiment' are troubling as well. I said I decided to roll the die 'a bunch of times'. Then I rolled the die 100 times. Did I decide in advance to carry out exactly 100 rolls? Or did I keep track of how 1 and 6 were doing, and decide to stop only when they pulled out in front? That would be worrisome, because if I were prepared to wait long enough, eventually 1 and 6 would be bound to come out on top. Our probability of roughly 1/15 was based on the assumption that the number of rolls is determined in advance.

So, OK, what if I had reported the experiment this way.

PROCEDURE: Roll the die 100 times; observe whether 1 and 6 are the two most frequent outcomes.

RESULTS: 1 and 6 were the two most frequent outcomes. According to the 'two-most-frequent-outcomes test', this would happen by chance only about 1 time in 15.

CONCLUSION: We reject the null hypothesis that the die is not loaded ($p < 1/15$).

DATA: All the data will be available on my website.

QUESTIONS: Does reporting the results this way convince you that I had decided beforehand how many times I would roll the die; what data I would collect; and what specific statistical tests I would apply to the data? Do you really believe that the data are now or will ever be available on my website? If the data aren't available and you complain about it, will anybody care?

When we see a published study, we tend to imagine that what the authors did to collect and analyze the data was something they decided upon in advance. In general, the authors will say nothing to contradict this, and their statistical analysis will be predicated on the assumption that this is in

fact the case. You believe this at your peril!

Authors very frequently state that the will share their data. Do not believe this! It is essentially never the case that you will be able to get the data.

But suppose—just suppose—that I really did decide in advance exactly what data I would collect, and what test I was going to do. And suppose I really do make all the data available. How does that change things?

Well, let's see. You could look at the data, and verify that 1 and 6 came up most frequently. Then you might be inclined to do a more sophisticated analysis of the data, say by computing a χ-squared statistic, and determining the probability of observing a value of this size by chance.

Whoops! You have just made a big mistake. The determination of the probability of observing an extreme value of some statistic is based on the assumption that you decide in advance what statistic you are going to use. Here you are deciding on the statistic after the fact. How would you feel if I did that? What if I decided after I did the experiment that just keeping track of the two most frequent outcomes wasn't a very powerful test, and that my experiment would be much more convincing if I used a χ-squared test:

PROCEDURE: Roll the die 100 times; compute the χ-squared statistic.

RESULTS: χ-squared statistic has such-and-such a value.

CONCLUSION: We reject the null hypothesis that the die is not loaded ($p < .01$).

DATA: All the data will be available on my website.

QUESTION: Now how do you feel?

ANSWER: Well, by now you are savvy enough to know that I didn't decide beforehand just what data I would collect, and what tests I would perform. So you aren't fooled by the notion that these results are something that would happen by chance at most one time in a hundred if the die were fair.

On the other hand, you may feel that the χ-squared statistic is the most natural statistic to compute in a case like this. If I were reporting some arcane statistic you'd never heard of before, you might suspect that I loaded my data into some statistics package, ran it through every possible test, and scanned through the results until I found a test with a publishable p-value. But you're not worried about that here, because I'm using the test you would have used, so it's just as if you did the experiment yourself.

Of course, there's still the problem that you don't know whether I decided

in advance how many times I would roll the die. This will make interpreting the p-value a bit tricky. You'll have to have some idea of the probability that the p-value will dip reasonably quickly down into the region $p < .01$ as I keep on rolling the die, keeping track of the χ-squared statistic as I go. But you probably have some gut feeling about this, or perhaps even some first-hand experience, and you sense that while it's not too hard to get p to dip down into the publishable region $p < .05$, getting it down to $p < .01$ is something you really can't hope to do more than maybe 1 time in 20. So it looks to you like this result really is significant. And once again, it really is just as if you had done the experiment yourself.

Or is it? What if, instead of extending my experiment until I got a good p-value, I kept starting it over? I might have done it once; got a bad χ-squared value; done it again; got another bad χ-squared value; and kept going until I got a really good value?

No, you know I would never have done something like that. That would be dishonest. Not in the same category as deciding as you go along when to stop collecting data, or deciding after the fact what tests to apply.

OK, let's now assume that this experiment was completely on the up-and-up: I decided in advance to roll the die 100 times and then compute the χ-squared statistic. I really do make all the raw data to you, and you survive the shock. Now is it just as if you had done the experiment yourself?

NO! It is *not* as if you had done the experiment yourself, because if you had done the experiment yourself, you would have known before you did it that you were going to do it. If you had known you were going to do the experiment, and what it was going to be, you could have determined in advance the probability of observing the various outcomes. And then, when the outcome was such that it would only happen by chance 1 time in a hundred, you would have been duly impressed.

But instead here I come, and say: 'Now, aren't you the lucky one? You could have done this experiment yourself, but you didn't have to, because I did it for you. You didn't even have to ask!' Are you going to go and analyze the data just as you would have done if you had planned and executed the experiment yourself? If you do, you'll say, 'Gee, how surprising! Just think of it: Only six billion people in the world, and only some small fraction of them style themselves scientists—say, 60 million at most. Here comes one of those 60 million, with an experiment I could have done myself only luckily I didn't have to, and the results are something that would only happen by chance one time in one hundred. That die *must* be loaded!'

I hope it will be clear why I don't believe you should be convinced by my experiment. I haven't put you in a position where it makes sense for you to reject the null hypothesis that all outcomes are equally likely, because you didn't do the experiment yourself.

In the next section, I'll describe what I could have done better. Meanwhile, I think I can easily convince you that the die really was loaded. I have only to reveal its approximate shape: $b \approx .8a$; $c \approx .65a$. Now your personal experience with cardboard boxes and other rectangular parallelopipeds of daily life will suggest that there should be a strong preference for 1 and 6. Learning that they were most frequent in 100 rolls will nail the question. In fact, I suspect you will find yourself wondering why the χ-squared test was only significant at the $p < .01$ level, and you will be relieved to know that I just made that part up. The actual results were such that there was no need to do a statistical test to know that the die was really loaded. That is, not if you were the one who did the experiment—which you weren't.

QUESTIONS: How sure are you that a die with proportions $(1, .8, .65)$ would generate a noticably unequal distribution when rolled 100 times? How would you go about trying to compute theoretical values for the various outcomes for a die of this shape? How would you compute a theoretical value for the probability that a cylindrical can will land on its side when tossed, as a function of the ratio of the can's height and radius?

What is to be done?

Suppose I undertake an investigation of the link between aspirin and batting averages, and I obtain what I regard to be significant results. Suppose that, to try to convey my excitement to you, I direct your attention to a permanent public archive, where you can find not only the data and results of my experiment, but a previously archived announcement I made of my intended experiment, laying out *beforehand* precisely what data I intended to collect, and precisely how I planned to analyze it. And suppose you are able to verify from this archive that what I said I was going to do was no more and no less than what I said I did. Now you are in a position that begins to approximate where you would be if you had done the study yourself.

This is not to say that you will wind up convinced of this putative link between aspirin and batting averages. You may feel that, while there could conceivably be a connection between aspirin and batting averages, the results I found are likely to have arisen by chance, even though my analysis

establishes to your satisfaction that the probability of this is as small as 1 in 100. You may find your suspicions growing when you look through this publicly searchable, publicly accessible, permanent, indelible archive, and discover that this particular study is one of a whole raft of studies that I or others you consider comparably reliable have registered, designed to test for possible links between aspirin and slugging percentage; aspirin and on-base percentage; aspirin and earned-run average; aspirin and bowling score; aspirin and GPA; aspirin and colon cancer; aspirin and academy awards; aspirin and ...—and there is no evidence that any of these other studies have amounted to a beanrow of anything. These other studies are just as deserving as this one to be treated as studies that you effectively 'did yourself without realizing it'. And of course when you do hundreds of studies, some will erroneously appear significant just by chance, even if the null hypothesis is true in every single case.

Of course, an even better way to win me over, and make it really as if I had done the study myself, would have been to attract my attention to the study as soon as you registered with the archive your intention of carrying it out. But it's too late for that, now, isn't it?

Now the interesting thing is that there is already a mechanism in place for registering in advance of a study what data are to be collected and how the data is to be analyzed. This mechanism is mandated by the federal government for any studies involving human subjects funded by the US Department of Health and Human Services, and separately, for studies of any drugs, medical devices, etc. regulated by the Food and Drug Administration. This means that investigators are already being forced to commit themselves in advance to how they are going to run their experiments. Well, at least in principle. Until we've had a look at a good sample of these applications, and taken some care to check how well what was planned agrees with what was actually done and reported, we may be inclined to doubt just how well this mechanism works.

But never mind that. The point is that experimenters are already forced to commit in advance to what they are planning to do. Now all we must do is try to figure out how to get them to divulge this information in advance.

So, who would want to lay out their experiments in advance, and why? Obviously, no one whose scientific career is predicated on fishing expeditions, or as we may now be tempted to call it, 'phishing', is going to want to have anything to do with a scheme like this. But reputable investigators have nothing to lose by making their intentions known in advance.

Archiving your plans in advance of the investigation would be a good way of signaling your scientific integrity. This could make it easier to get funding, get results published, etc.

If I were a funding agency, or a reviewer for a funding agency, I would look kindly on investigators who have an established policy of archiving the plans of their investigations in advance. Such a policy would impress me as a sign of honorable intent, and just as important, it would give me a way to check over their track record. Turning from the investigator to the particular investigation, I would look kindly on proposals whose study plans were either already archived, or (more likely) where there was a commitment to archive the plans before the start of the study. In fact, if I were in a position to do so, I would make funding contingent on pre-archiving any study paid for in whole or in part with the funds being allocated.

Other parties who could benefit from a pre-archiving scheme would be the journals where results of studies are published. If I were a medical journal, I would look kindly on studies that were archived in advance. In fact, if I were an editor, I would consider for publication only studies archived in advance, and if I were a reviewer, I would consent to review only studies that had been archived in advance.

Using the word 'archive' makes this all sound too formal, like another layer of red tape added to what is already an onerous process. Really it could be as simple as inaugurating a website, and inviting investigators to contribute study plans beforehand, and results afterwards. In the case of studies involving human subjects, they will already have had to complete these study plans, and of course the will be compiling final reports for publication. All it would take would be a few mouse clicks to make these study plans permanently, publicly accessible.

Other benefits of this proposal will suggest themselves. With investigators announcing their plans beforehand, in a publicly searchable forum, there would be an opportunity for others to scrutinize the plans, suggest possible improvements to study design and/or statistical technique (to be adopted before the study begins!).

In fact, one can imagine the development of 'contingent studies'. I may examine your study plan; discover possible improvements to the statistical analysis; and announce beforehand my intention of submitting your results to an alternative analysis.

One can also imagine that such public scrutiny might not appeal to certain parties. Not everyone is ready for the rough-and-tumble of free debate.

Someone might steal my research plan, or swipe my data and make better use of it than I. Such parties could be given the option of having the precise details of their study kept secret until the study is completed, revealing only the general nature of the study (so that it will still be possible to determine just how many aspirin-and-batting-average studies are in the works). I don't see why any funding agency would want to allow this kind of nonsense, but we want to maximize the incentive for individual researchers to take the plunge and start doing at least an approximation real science.

Our motto (cribbed from Garrison Keillor: 'REAL SCIENCE: Why not pretty soon?'