# Supreme Court rules against overreliance on $p$-values

Ben Klemens

8 April 2011

This is the case of Matrixx Initiatives Inc. et al. vs. James Siracusano et al. (PDF Opinion[1])

The question in the case is whether Matrixx responded correctly to a doctor's publishe findings regarding ten cases of people out in the public who used their flagship product, Zicam, and then permanently lost their sense of smell. If we were running a controlled experiment, ten cases out of tens of thousands is not statistically significant. Matrixx is a publicly traded company, so it is their obligation to reveal to shareholders all pertinent info, but Matrixx didn't disclose the news about this study, because the results were not statistically significant.

Initially, Matrixx did a ham-fisted job of responding: they sent a cease-and-desist letter to the author of the paper telling him that he did not have permission to use the brand name Zicam in his paper, which just made them look like bullies, created a paper trail that they had seen the study, and which was irrelevant anyway, because tradmark $\neq$ copyright, and you don't need any permission from anybody to make true and above-the-board statements about a product by name. You think the Chicago Tribune[2] or Forbes[3] asked for permission before repeatedly using the word Zicam in their coverage? But enough about what looks like a solid botch of intellectual property law.

Let's get back to the botching of statistics. The key claim that Justice Sotomayor spent the ruling tearing apart was that "reports that do not reveal a statistically significant increased risk of adverse events from product use are not material information." That, is Matrixx claimed a bright-line rule that if a study turns up $p > 0.05$, then it is immaterial.

I won't go into great detail on the Court's argument, because I'm writing on a statistics and computing blog, and I do not believe that any of you reading this blog would take a bright-line $p$-value rule at all seriously in your own work. You can maybe find some stats textbooks that suggest something like this to undergrads, but I'd guess that the authors feel terrible about oversimplifying so much. You may believe that a journal has a bright-line editorial custom of only publishing studies that eke out a $p < 0.05$, but at the same time make nasty comments about how the system is broken.

[1] `http://www.supremecourt.gov/opinions/10pdf/09-1156.pdf`

[2] `http://www.chicagotribune.com/news/nationworld/sc-dc-0323-court-business-20110322,0,4462213.story`

[3] `http://blogs.forbes.com/billsinger/2011/04/01/buffett-sokol-zicam-matrixx-supreme-court/`

Like neckties, it's one of those self-perpetuating customs that we all know we'd be better off without.

The Court's discussion begins at *A* on page nine of the PDF linked at the top of this column, and I give you the page number because it is recommende reading. I worked in tech law (until it got boring), and the Supreme Court rulings were always the funnest part of the work. First, the ruling is about a specific question, which may not be what the press yammers about; you may be surprised that the case is really about a legal technicality, and that the Court really wants to say something else but instead winds up writing a ruling that just keeps some detail of the legal machinery clean. The case of Westboro Baptist Church hurling hompohobic invective at a soldier's funeral (Opinion PDF[4]) made mention here and there of speech which is offensive and onerous ("Because this Nation has chosen to protect even hurtful speech [. . . ], Westboro must be shielded from tort liability for its picketing. . . "), but the legal logic is entirely about who had obtained what permits when and where people were standing. There, the subtext itself makes for good reading.

Because these are typically rulings about the Big Questions, like whether we can we derive cetainty out of studies rooted in probabilites, so they are much more readable than the average opinion (especially once you get into the habit of just letting the excess of citations and footnotes wash past you). So I encourage you to see how a lawyer tears apart somebody's claim that $p$-values provide a bright-line test for evidence's relevance. Pay especial attention to footnote six, in which Justice Sotomayor defines what a $p$-value is. I wish I was writing another textbook so I could cite the Supreme Court on this.

The justices instead reiterated a prior ruling that something needs to be disclosed to investors if there is "a substantial likelihood that the disclosure of the omitted fact would have been viewed by the reasonable investor as having significantly altered the 'total mix' of information made available." If you're the sort of person who thinks in terms of Frequentists vs Bayesians, that means you're a Bayesian, and that probably means that you're salivating right now, because the Supreme Court just ruled that information is relevant to the extent that it causes a reasonable person to update his or her subjective prior.

**The right null for the job** Following a common pattern in the medical literature, there is anecdotal evidence that Zicam caused a burning sensation followed by a loss of smell, backed up by some prior knowledge that zinc has been known to have deleterious effects on certain types of tissue. There's a small-$n$ problem at the core of this: if one in ten thousand suffer an effect, then clinical trials of a hundred patients have no chance of passing the bright-line of $p < 0.05$, but after a million people use it, then we expect a hundred people will have suffered a permanent loss of their sense of smell.

The null hypothesis in a study is typically of the form *nothing happened, there are no differences, nothing of significance is going on*. This is a good default because your typical researcher is running a study because he or she really believes that there's something going on, and so *nothing happened* correctly sets the bar high.

[4] `www.supremecourt.gov/opinions/10pdf/09-751.pdf`

For the medical literature, when it is asking whether harm is caused, this is not a helpful null hypothesis. Say that a study's null is that Drugacil does no harm, but the data finds that Drugacil kills people, with $p = 0.75$. There's a 75% likelihood that the thing about killing people was just random noise, and a skeptical researcher might retain the belief that nothing happened until given convincing evidence that something did, but I sure ain't using Drugacil. The correct null here is that harm was caused, and in an ideal world we reject it only when we are confident that there is no harm.

This isn't to say that all evidence is relevant evidence, and other inquiries in other contexts will play out differently. There's still the micronumerosity problem, potential ethical issues of such a study, et cetera. But this point is worth adding to the Supremes' already long list of explanations for why a bright-line $p$-value test doesn't work: sometimes the right null hypothesis shouldn't be that nothing happened, and sometimes, evidence that might be due to chance is still important and in need of consideration.

Why are there still all those undergrad textbooks that push for a bright-line $p$-value test? Because we want certainty. We don't want to live in a world where statistics only speaks in probabilities and where context always matters. But here we are.