Quantamagazine

A Revealer of Secrets in the Data of Life and the Universe

The statistician Donald Richards lives to uncover subtle patterns hiding in real-world data.

By Natalie Wolchover

Jessica Kourkounis for Quanta Magazine

Donald Richards at Pennsylvania State University, where he is a professor of statistics, in February.

In statistics, abstract math meets real life. To find meaning in unruly sets of raw numbers, statisticians like Donald Richards first look for associations: statistical links between, say, smoking and lung cancer, or the closing values of the New York Stock Exchange one day and the Tokyo exchange the next. Further study can then probe whether one phenomenon causes the other, or if both have common causes.

"Statistics is a way of analyzing data and discovering the inner hidden secrets being concealed by the data," Richards said over Skype from his home in Pennsylvania in January. "Can we find patterns that tell us that climate change is underway? Can we find patterns that suggest that bitcoin has topped out? That's what we're constantly searching for — patterns."

The patterns can be subtle. However, the search for them is not esoteric, in his view, but rather "the only thing that anybody with brains should be doing with their life."

In the lilt of his native Jamaica, Richards, 63, describes statisticians as innovators, ever in search of new mathematical tools for finding hidden associations between phenomena, and thus furthering the ancient quest to link causes and effects. How, for example, did people first figure out what they could eat? "In Jamaica there's a tree called the ackee tree," he said. "When the ackee fruit is not ripe, it's highly poisonous, but [as a deadly search for correlations must have revealed] when it's ripe, if cleaned properly it can be cooked and eaten."

Throughout a career that has taken him to the universities of the West Indies, North Carolina, Wyoming, Virginia and Pennsylvania State University in State College, where he is currently a professor, Richards has derived many new mathematical formulas for use in statistics. He has also applied them to sleuth out correlations in real data, such as galaxy surveys, financial derivatives, agricultural data and the Affordable Care Act insurance market. A formula that he proved with his mentor and longtime collaborator, Kenneth Gross, in the 1980s assisted in the development of cellphone data transmission protocols that made calls statistically less likely to drop. "Every time I see someone using a cellphone I wonder if I can get them to fork over just one cent," he joked.

Lately, Richards has taken up a powerful new tool for detecting correlations, the "distance correlation" method. In 2014, he and his wife, the Jamaican-American astrophysicist Mercedes Richards, and a third coauthor applied the distance correlation method to detect previously unknown associations among 63,500 galaxies – identifying, for instance, that some were "starburst galaxies" that quickly burn themselves up.

Real life also has a way of intruding in unwelcome ways. Mercedes died in 2016 at the age of 60, and Gross died last year at 78. Richards wears his grief on his sleeve, even as he revels in the joy of statistics. After the January interview, *Quanta* followed up with Richards by phone in March. The following questions and answers have been edited and condensed for clarity.

Growing up in Jamaica, did you always enjoy math?

My mother was very good at mathematics. Because she was a teacher, learning did not stop simply because school was out. During the summer her three sons, and even our friends from neighboring homes, if they wanted to play with us, were required to do schoolwork at our house from 9 until 12. After lunch, the rest of the day was ours.

One summer, when I was 10, she had me work through chapters of one of her old algebra books, *Higher Algebra* by Hall and Knight. I didn't understand it; it was just silly algebra. I had no reason to think it was useful for anything. But then the next summer, she said, "Why don't you start with Hall and Knight again." And the second time those chapters were a lot easier to handle.

In high school, at Kingston College, I hit the jackpot with a superb math teacher in third form [the equivalent of ninth grade]. In fourth form, I hit the jackpot again as I had another superb math teacher. Then in fifth form, I hit the Powerball; I had a superb math teacher who taught me for three years straight [Kingston goes up to grade 13]. So coming out of high school I had one of the best preparations possible for a math student in Jamaica, and I would dare say anywhere in the British Commonwealth, and maybe even in the British Empire.

Yet you were the only student in your math cohort to pursue it academically. Why is that?

There was a tendency in Jamaica for people to shy away from mathematics because of the economic considerations. Incomes would tend to be substantially smaller than if you were to become, say, a doctor, lawyer or engineer.

Jessica Kourkounis for Quanta Magazine

How did you find your way to statistics?

When I went to the University of the West Indies, mathematics was everything. If I had tried to study physics or chemistry or biology or law or medicine, I would have been a complete failure. I took a course on probability and statistics in my second year from Rameshwar Gupta, who eventually became my thesis adviser. He was an expert in multivariate analysis [which concerns the relationships between many variables], and so that became my field. I'm a great believer in working with people I like. And I could also see that there were lots of interesting mathematics problems working with him, so it was the best of both worlds.

That's how I got into statistics, but I didn't really become a statistician until I was fairly close to getting tenure at the University of North Carolina. I knew all the formulas, but the art and joy of analyzing data did not really hit home until then. Before that, I was basically just a mathematician pretending to be a statistician.

What’s the difference between a mathematician and a statistician?

It’s so hard to decide where the boundaries lie, but I would say that mathematics is a field that specializes in deductive logic: They lay down a bunch of axioms and then they try to deduce the logical consequences of those axioms. Two points define a line, and so on and so forth, and then off you go. Whereas statistics is more the art of inductive logic: We look at the end result and we try to understand what could have caused such an end result.

We “mathematical statisticians” come up with formulas that help with the inductive aspect of our data analyses. Let me try and give you a simple example: Suppose I have a coin in my hand and I ask, is the coin a fair coin or is it biased toward heads or tails? What we would do is toss the coin 100 times, and if the percentage of heads is too far from one-half in either direction then we would be inclined to say that the coin seems to be unfair. But the question is, how far is too far? Suppose we got 48 percent heads. How about 40 percent heads? The question of how far from one-half is too far is answered by using various mathematical formulas and probabilistic calculations.

So, how far is too far? I would guess 40.

You made a very good guess! I am very impressed! This question has consumed people’s attention for centuries. My cardinal rules are: If a coin is tossed 100 times and the number of heads is zero to 40, or 60 to 100, then you have strong statistical evidence that the coin is biased. Moreover, depending on how much money is at stake, if you get 41 or 59 heads, then you’re well-advised to leave the casino immediately. Even at 42 or 58 heads, you should be nervous about the tosser.

When people come to you for advice, you tell them to stick to the middle road in most endeavors. What do you mean by that?

Jessica Kourkounis for Quanta Magazine

You go safest in the middle course. My advice to go with the middle course is taken from Benjamin Graham's book, *The Intelligent Investor*, a book whose acolytes generally do very well in the financial markets. If people know nothing about financial matters, it would be my advice for them to go with large mutual funds that are widely diversified and do very little trading. Not trying to make tons of money overnight. Don't buy bitcoin. Don't buy anything that's newfangled, by the way. There are people who make huge amounts of money buying newfangled stuff, but the bulk of people who do that eventually go broke.

People, even some statisticians, might be surprised to learn there's a groundbreaking new mathematical tool for finding statistical relationships — the distance correlation method.

Without a doubt there have been innovations. Let me explain the situation regarding the concept of correlation. People like Karl Pearson and Francis Galton in the late 1800s pioneered the concept of what we now call the Pearson correlation coefficient, the widely used measure of correlation.

Galton studied the correlation between heights of fathers and heights of sons, right?

Yes. Galton was trying to understand how to grow a wonderful human body — bigger, taller, stronger, faster, more intelligent. And so he took a large collection of fathers and sons and plotted the data — you can find the scatter plots on the internet, actually — and he fit a straight line through the data, finding that if your father is tall then you are likely to be tall. This method, which Karl Pearson formalized in what's now known as the Pearson correlation coefficient, has its advantages. But it also has flaws. For example, it can be applied only when you're trying to study the relationship between one-dimensional variables — between height and weight, or height of father and height of son. If you want to study the correlation between one batch of variables and another batch, then there is no single Pearson correlation to measure the strength of an association. A second problem, which people often overlook in everyday applications, is that the Pearson correlation coefficient should be used only when there is a reasonably linear relationship between the two variables. If the relationship is highly nonlinear then this method is inapplicable.

What's an example of a nonlinear relationship between variables?

Height and age. Imagine that I take a human being and I measure their height from birth to adulthood. It's an S-shaped curve, right? A linear relationship does not apply.

It seems like there might be more cases like that than linear relationships.

I guarantee you that the bulk of applications of the Pearson correlation coefficient are invalid. Often it's being used where it's really not helpful, and leading to conclusions that are not supported by the data. Health data, financial data, astrophysical data, meteorological data: People just jump right in and start using Pearson's correlation without any thought to whether there is any plausibly linear relationship between the variables.

We've been aware for a long, long time of the need for a correlation coefficient that can be applied regardless of whether there is a linear or nonlinear relationship, and can be applied when we have batches of variables on both sides. And that's where the distance correlation came in. Around 2007, Gábor Székely, Maria Rizzo and Nail Bakirov wrote these papers in which they defined the concept of distance correlation and applied it to data. And when I saw those papers I said, "Aha! I know how to apply this to astrophysical data." And it was exactly to the data that my wife, Mercedes, Elizabeth Martínez-Gómez [of the Autonomous Technology Institute of Mexico] and I had been thinking about.

Courtesy of Donald Richards

Donald and Mercedes Richards in 2015 during a visit to the Heidelberg Institute for Theoretical Studies in Germany.

Are you continuing your studies of astrophysical data?

Mercedes and I were going to write a third paper on distance correlation for other astrophysical data. I have the notes on that, and I need to do it. But as you can imagine, every time I look at it I feel a little bit… well, the word that comes to mind is "depressed." But that's something I really should finish. Ken Gross and I were going to write another paper on what's called the hypergeometric functions of matrix argument. We worked on it for 20 years; I have hundreds of pages of notes. That paper will be over 100 pages long, and I need to sit down and do it.

At this very moment, two coauthors and I are revising a manuscript that will provide applications of distance correlation to manufacturing processes. We have found ways of applying the method to determine whether processes are in control or not, meaning the average number of defects per product is below a certain threshold.

How does distance correlation work?

This is where the concept of a Fourier transform comes in. A Fourier transform is a way of breaking up a mathematical function into its component frequencies, similar to how a music chord can be decomposed into its constituent notes. All functions can be uniquely characterized by Fourier transforms, so people started to try to define the concept of a measure of correlation by using Fourier transforms. If you give me two probability distributions — the statistical spread of values that a variable takes on — and if I want to test whether the two distributions are the same, all I have to do is calculate their Fourier transforms. If these are equal then I know that the two probability distributions had to be equal to begin with. The distance correlation coefficient, in layman's terms, is a measure of how far apart these Fourier transforms are.

You wrote a paper last year giving examples where distance correlation improves on Pearson's method. Talk about the case of homicide rates and state guns laws.

This was prompted by an opinion piece in *The Washington Post* in 2015, by Eugene Volokh, a professor of law at UCLA. The title of the article is "Zero Correlation Between State Homicide Rate and State Gun Laws." What he did was — you know, my eyes bugged out; I couldn't believe it — he found some data on the states' Brady scores, which are ratings based on the toughness of their gun laws, and he plotted the Brady scores on an *x*-*y* plot against the homicide rates in each of these states. And if you look at the plot, it looks like there's no pattern. He used Excel or something to fit a straight line to this data set, and he calculated the Pearson correlation coefficient for this data set, and it came out to be nearly zero. And he said, "Aha, zero correlation between state homicide rate and state gun laws."

That's not kosher?

I was horrified. There are so many things wrong with this analysis. The first thing you notice in the scatter plot is that there's one dot which is way, way out from the others, with both a high Brady score and high homicide rate. That turns out to be the District of Columbia, which is not a state; it's really a city, so if you include it in the analysis, because it's so far away from everybody else, it's going to have a major effect on the slope of the regression line. That's the first complaint; he should have removed that data point — you learn that in Stat 100. If you remove it and refit the linear regression line, the Pearson correlation is not zero, actually.

But should you even fit a linear regression line to this data set? If you look at the rest of the data, you don't see any linearity to the relationship, and it's easy to understand why: There are bunches of points that correspond to geographic and culturally similar regions. If you break up the states by region, then you see reasonably linear relationships starting to show up in the scatter plots. And then in each case, you find that the higher the Brady score, the lower the homicide rate.

Distance correlation does an even better job without having to split things up, right?

Exactly. My wife and I did these calculations in the fall of '15, when we saw the opinion piece. She was a whiz at computer programming — I miss her just for that! It's a silly reason to miss her, don't you think? She calculated the distance correlation coefficient. And even with the District of Columbia included, it signaled that there was a statistically significant relationship between Brady scores and homicide rates.

There are so many examples like that of statistics getting misused in public discourse. I wonder whether you spend time thinking about how to improve the use of statistics by nonexperts.

Yes, there are these horrifying misapplications of statistics. And I have to tell you the truth: Sometimes I just give up. It just never stops, the number of misapplications. As an educator I know I have a duty to continually educate the world. But I'm just one little guy. At some point, you know what, I'm going to go cook myself a nice dinner and have a glass of wine and let the world do whatever it wants.

Do you hope that by developing better tools — like distance correlation — that eventually these methods will seep out into more common use?

Yes, I hope so. And in fact, I have heard that one of the big pharmaceutical companies is now starting to use distance correlation methods. And I know that people in academia are using it more. I hope to live long enough to see distance correlation be a standard pulldown tab in Excel, or if not Excel, certainly on Wolfram Alpha. You enter your *x*-*y* data, and boom: It gives you the distance correlation. I live for that day!

This article was reprinted on TheAtlantic.com.