PERMANENT LINK | MARCH 23, 2015

Campbell Harvey on Randomness, Skill, and Investment Strategies
EconTalk Episode with Campbell Harvey
Hosted by Russ Roberts

Continuing Education... Paul R...

Home | EconTalk | Archives | Permanent Link and Comments

Campbell Harvey of Duke University talks with EconTalk host Russ Roberts about his research evaluating various investment and trading strategies and the challenge of measuring their effectiveness. Topics discussed include skill vs. luck, self-deception, the measures of statistical significance, skewness in investment returns, and the potential of big data.

How do I listen to a podcast? Right-click or Option-click, and select “Save Link/Target As MP3.”

Related Readings

About this week’s guest:

- Campbell Harvey’s Home page
- Garden of Econ. Campbell Harvey’s Blog.

About ideas and people mentioned in this podcast episode:

Articles:


Web Pages and Resources:

- “Significant”. XKCD comic. Do jelly beans cause acne?
- Sharpe Ratio. Investopedia.
- Warren Buffett. Wikipedia.
- Search for the Higgs boson. Wikipedia.

Podcast Episodes, Videos, and Blog Entries:

- EconTalk episodes with Nassim Taleb. EconTalk.

Highlights
0:33 Intro. [Recording date: March 6, 2015.] **Russ:** Our topic for today is in some sense randomness, one of the deep ideas in thinking about complexity and causality. And a jumping off point, though, we’re going to use a recent paper you wrote with Van Liu, “Evaluating Trading Strategies,” which was published in the *Journal of Portfolio Management.* And we may get into some additional issues along the way. Let’s start by reviewing the standard way that we evaluate statistical significance in economics for example or other applications of regression analysis. You’ll hear people talk about a t-statistic being greater than 2. And what does that represent? What are we trying to measure there? What are we trying to assess when we make a claim about significance of, say, one variable on another? **Guest:** So, the usual procedure—we actually think about trying to minimize the chance that a finding is actually a fluke. And it comes down to a concept called the p-value or probability value. And what we usually try to do is to have 95% confidence that the finding is actually a true finding. And by definition, there’s a 5% chance that the finding is a fluke. And when you do that in standard sort of statistic analysis, that is a so-called ‘two sigma’ (2-σ) type of rule. And often this is quoted popularly, in surveys and things like that, [,7], a confidence level of plus or minus a percent. And that’s the same 95% confidence interval that leads to this two sigma rule which is the same thing as a t-statistic of 2. **Russ:** And this is a convention in economics, that 2 standard deviations, two sigma is therefore probably not a fluke. The 95% level of significance. And I want to add one other important point before we go on: when we talk about significance, all we mean in this technical conversation is ‘different from random’: that there is some relationship. It doesn’t mean what it means in everyday language, which means important. So a finding can show a relationship between two variables that’s significant but quite small. So it’s significant statistically by Guest: Yeah. In real life. **Russ:** There’s two different concepts and both of them are important. We’re talking about **statistical significance** by a two sigma rule. There’s another concept that’s equally important called **economic significance:** is this fact really a big deal or is it small in terms of the big picture of things? **Russ:** So, as I said, it’s a convention that 95% means, Well, there’s only a 5% chance. And for many people that sounds--and many economists accept, that that’s like, well, if it’s only 1 in 20 then it’s probably real. We’ve ruled out the likelihood that this is just a fluke. But as you argue in your paper--and we’re going to talk about some different examples of this--when the number of tests that we’re making starts to increase, that statistical technique is not as convincing. So, to set that up I’d like you to talk about the Higgs boson. Which seems far away from finance, but I found that to be a fascinating example to help us think about things. **Guest:** Yes, certainly. So the Higgs discovery was complicated. It was complicated for many reasons. One, they had to build a collider that cost $5 billion to construct. But once it was constructed, they knew what they needed to find. And this was a particular decay signature that would be consistent with the Higgs boson. But the problem was that the same signature could arise just by random chance. And the number of collisions that they were doing and signatures that were being yielded was on the order of 5 trillion. So, just a huge number of possible false findings for the Higgs. **Russ:** And we’re looking for--we are trying to identify the Higgs—a particular subatomic particle. **Guest:** Exactly. So, what they had to do was, given the extreme number of tests, they had to have a very different sort of cutoff for establishing a discovery or establishing statistical significance. And instead of using the two sigma rule, they used a five sigma rule. So, way different from what we’re used to. And this reflects just the number of tests that were actually being conducted. **Russ:** So the idea--try to give me the intuition of this. I’m going to collide a lot of things--I’m going to collide particles many, many times, trillions of times. And we know that’s going to generate lots of false positives, decay signatures that look like the Higgs but are not. Correct? **Guest:** That is correct. So you have to be really[?] sure. **Russ:** So, shouldn’t I just--isn’t the ‘really sureness’ just the fact that this is easily confused rather than--what do you mean by the number of tests? **Guest:** Well, really what we’re talking about in the Higgs example are the number of collisions that are taking place. And I’m simplifying what they actually did at the collider. There are many different tests actually going on. But the fact is that sometimes you would get a signature that looked like the Higgs but really wasn’t the Higgs. So, in order to actually--and nobody had actually discovered the Higgs. This was the first opportunity. So, they had to be really sure that they were not being fooled by the random sort of occurrences of something that looked like the Higgs. So, to do this they had to be, as I say, five-sigma confident that it really existed. **Russ:** Now, it’s a little bit, for non-statisticians, 5 versus 2, is actually a little less. Right? That’s just a misleading. So, it’s a little more than twice as big; so we’re requiring the result to be a little more comfortable. But as we move numbers of sigmas away from zero, it’s a much smaller chance than say a little more than twice as likely that it’s by chance? Right? **Guest:** Yeah. You mentioned the 5% or the 95% confidence. That means that 1 out of 20 will be a fluke. So, for a 5-sigma, it’s 1 divided by 1.5 million. So, that is a very small-- **Russ:** But it could still be a fluke? **Guest:** Yes, it could be. **Russ:** It’s a weird thing, because you’d think you either see it or you don’t. But I guess it’s elusive, and there are things that look like it but aren’t it, is what you are really saying.
**Guest:** Yeah. So can I give another example that I think is going to further intuition? **Russ:** Sure. **Guest:** It’s the famous Jellybean comic. Have you seen that before? **Russ:** I have, and we’ll put a link up to that. It’s one of my all-time favorites. So, yeah, describe it. **Guest:** So, this is a famous cartoon called “Significant.” **Russ:** It’s XKCD—the cartoon is in a series. **Guest:** Exactly. So, somebody makes a statement: I think that jelly beans cause acne. So, they said, okay, scientists: Go investigate. So the scientists go and do a trial. And the trial would involve I guess giving some people jelly beans and other people without the jelly beans. And then they would test to see if there’s any significant difference, I guess, the number of pimples for people that took the jelly beans and ones that didn’t. And the test comes back, and there’s no difference. There’s no significant difference between the two. So basically the next frame of the comic is, well, maybe it’s not the jelly bean itself, but the color of the jelly bean. So, then, the comic goes and the scientists test different colors of jelly beans. So, again, the trial would be, let’s say a group of people get some red jelly beans and others don’t get any jelly beans. And they go through all of the colors. So, red, there’s no effects, there’s no difference in the amount of acne. And orange, yellow, purple, brown, black, ... **Russ:** Fuchsia, mauve. **Guest:** Exactly. The 20th test is green. And they find that there’s a relation with the green. So they declare that green jelly beans cause acne, and that’s actually gets into the headlines of the media the next day: Green Jelly Beans Cause Acne. **Russ:** With a 95% chance that it’s not random. That is, only a 5% chance that it’s random. **Guest:** That is true. So that’s what the significance means in this particular case. So, everybody knows that there shouldn’t be a significant effect, because it doesn’t make any sense, what they’re actually doing; the original test was the correct test: jelly beans versus no jelly beans. But, the more tests that you actually do, it’s possible to get a fluke. It’s something random. **Russ:** And if you do 20, you expect one of them to, by chance, show that relationship. **Guest:** That’s right. So that’s why, when you do 20 tests, you can’t use the 2-sigma rule. So by the 2-sigma rule, if you try 20 things, then the odds are that something is going to come up as a fluke, as a finding that really isn’t a true finding. So, if you are going one test—as the original test that they did, in the comic, were they tested on a group of people; they gave them jelly beans and the other group, no jelly beans. In that test, two sigma is fine. That’s a single test. But once you start doing multiple tests then you run the risk that something is going to show up as a fluke, and two sigma is not good enough.

**Russ:** So, let’s take the example you give in your paper, which is really beautifully done. And although there are some technical things in the paper, I think the average person can get the idea of what you’ve done there, which is: You present, at the beginning, a particular trading strategy. Meaning a way to “beat the stock market” and make a lot of money. And then the strategy that you show, you of course tested over a long period of time, because people know that, in a short test, maybe by luck you would just do well. But it’s over many years. And although it doesn’t do so great in the first year, it then does very, very well consistently, including through the financial crisis of 2008, when many people lost their shirts and other pieces of clothing. And it looks like a fantastic strategy. And you evaluate that with the Sharpe Ratio. And then in general terms if you can about what the Sharpe Ratio is trying to measure as a way of evaluating in particular stock trading—in general terms. **Guest:** So, the Sharpe Ratio is basically the excess return on the strategy. Just think of it as the average annual return on the strategy divided by the volatility of the strategy. **Russ:** And, indeed, there’s a direct link, a direct relationship between the Sharpe Ratio and the t-statistic that we were just talking about. So, they are mathematically linked, and a high Sharpe Ratio means a high t-statistic. Which means that the strategy is a strategy that generates a return that is significantly greater than zero. **Russ:** And that’s actually relative to a so-called risk-free return. **Guest:** Yes. Usually you subtract out a benchmark, so just a risk-free. **Russ:** And I might want to be comparing it to, say, a different benchmark. Right? Say an index mutual fund, which a lot of people hold. I’m often interested in, Did this strategy, this manager or this different kind of investment pattern, did it outperform the S&P 500 (Standard and Poor’s 500-stock index)? That’s not risk free but it’s relatively cheap, low cost, because it’s automated. Is that correct? **Guest:** That is correct. So it’s often used to look at excess performance. That is a strategy: you can think of investing in somebody who’s got a certain return stream and then think of it as shorting the S&P 500 futures. And that’s a strategy on its own. And the question is: Do you get a return, an average return on that strategy, that is significantly different from zero? **Russ:** Say that again? **Guest:** Is significantly different from zero. So, that’s basically, if it’s different from zero—if it’s above zero, that is an indication of skill: that the strategy actually has something that the market doesn’t recognize and leads to some positive return, on average. And that’s what we all seek. We seek to beat the market. **Russ:** I’d say it in a different way. We are very—it’s easy to be seduced. We do seek it, but we also, we desperately would love to have sort of that inside path, the secret strategy: ’The suckers, they’re just, they’re accepting that mediocre return; but I’ve got the genius advisor running my money, giving me financial advice, and so I’m making a premium.’ We have a real urge to have that. **Guest:** Yes. We definitely want to be better than the average. And this is a prime target. And we want to allocate our money to
managers that we believe are skilled. And skill means that you can outperform the market. **Russ: And one of our lessons today, for me, thinking about these issues is how difficult it is to measure skill. So let's--since we're talking about this strategy which you start off the paper with, it has a significant Sharpe Ratio, right? It's much better than the average return. **Guest: It looks really good. You mention that there was a bit of a downturn in the first year, but it wasn't really that much--like 4%--and then it performs really well, all the way through the end of the sample. You look at it--it is significant. It is like 2.5-sigma. So it means that with the usual statistical test, you would expect this strategy to be true. And that this is something that actually does beat the market. **Russ: And here I am, naively investing my portfolio in a lot of index funds, and I obviously should switch. I'm losing money. I'm a fool, because I should be doing this. **Guest: That's what it looks like. **Russ: Yeah. But explain how you generated that fabulous strategy and why it's a bad idea. Or at least not significantly proven to be a good idea; standard deviations above the likelihood that it's random. **Guest: Sure. So the opening panel of my paper shows this great strategy. Very impressive, 2.5-sigma strategy. The second panel of the paper shows that strategy, plus 199 other strategies. And it turns out that what I did was I generated random numbers. **Russ: Say that again? I'm sorry, you cut out there for me, anyway. **Guest: I generated random numbers. Basically it's completely--there's no real data. I generated a series of random numbers, with an average return of zero and a volatility that mimicked the S&P 500. And on this graph, I plot the cumulative returns of these 200 strategies. And you can see that on average, the 200 deliver about a zero return over the 10-year period. But on the tails, you can see the original strategy that I presented, that had a 2.5 sigma, that did really well. And you can see on the other side the worst strategy, which had a 2.5 sigma below zero. So basically, what appeared to be a great strategy, was purely generated by random numbers, had nothing to do with beating the market. And again, this is a situation where you've got 200 random strategies. Some of them are going to look significant when they are not significant. Every single one of these strategies by definition had zero skill. Because I fixed the return, when I'm simulating the numbers, to have an average of zero.

---

**Russ: Let's do one more example, then we'll get to what the implications are. So, the other example you give, one of my favorites, is--I'm going to use the football example. I get an email from a football predictor who says, 'I know who is going to win Monday night. I know which team you should bet on for Monday night football.' And I get this email, and I think, well, these guys are just a bunch of hacks. I'm not going to pay any attention to it. But it turns out to be right; and of course who knows? It's a 50-50 chance. But then, for the next 10 weeks he keeps sending me the picks, and I happen to notice that for 10 weeks in a row he gets it right every time. And I know that that can't be done by chance, 10 picks in a row. He must be a genius. And I'm a sucker. Why? **Guest: Yes. So this is a classic example. So let me set up what actually happens. So, let's say after those 10 weeks in a row you actually subscribe to this person's predictions. And then they don't do so well, after the 10 weeks. And the reason is that the original strategy was basically: Send an email to 100,000 people, and in 50,000 of those emails you say that Team A is going to win on Monday. And in 50,000 you say Team B is going to win on Monday. And then, if Team A wins, the next week you only send to the people that got the correct prediction. So, the next week you do the same thing. 25,000 for Team A, 25,000 for Team B. And you continue doing this. And the size of the number of emails decreases every single week, until after that 10th week, there are 97 people that got 10 picks in a row correct. So you harvest 97 suckers out of this. **Russ: Who are willing to pay a huge amount of money, because you've got inside information, obviously. And I can make a fortune, all on your recommendation. **Guest: That's what it looks like. And the fact is, that this is basically a strategy of no skill. Basically, 50-50 every single week. There's no skill whatsoever. But it looks like skill. So, again, when you realize what is actually going on, you can't use the same sort of statistical significance. Because, in the usual case, to get 10 in a row, that is highly significant. But, given what you know has happened, it can't be significant. It's exactly what you expect. **Russ: So, that leads us to the deep question: Is Warren Buffett a smart man? I mean, he is called the 'Sage of Omaha.' He's done very, very well. He makes a lot of money relative to his competitors. Berkshire Hathaway, which is his stock version of his portfolio, is a wild success. So obviously he's a genius. True or false? **Guest: So, I've not actually studied the Berkshire Hathaway data. So, I'm not going to make a judgment on it. But maybe I will. It's a good idea for my research. So, this is--when you've got 10,000 or more managers that people are going to look really good, year after year, just because you've got over 10,000 managers, purely by chance. So, they could be monkeys throwing darts at the Wall Street Journal. So, this is exactly the same situation that I'm highlighting in my paper. That many managers will, potentially, for 10 years in a row, beat their benchmark. And it will be a result of luck, that there will not be any skill. And this is also important: That you could have many managers that are skilled, that are excellent managers-- **Russ: Yeah, the flip side. The flip side is hard to remember. **Guest: Yeah. This is very important. Because often your manager doesn't perform as well as you wanted, for, let's say, 2 or 3 years in a row. And then you ditch that manager. And that's a mistake. Because it's possible that that manager is a skilled manager and basically suffered from some bad luck. It's the randomness that can get you. So, mistakes are made on
Russ: So, this--it raises a very tough question, and it's a question that's pervasive. I may bring up some other sports examples in a minute. But just in case I don't, you know, people will talk about so-and-so is the greatest coach of all time; so-and-so is the greatest quarterback of all time. Or that Campbell Harvey is the best finance professor of all time. Because of a variety of factors. And of course there's a random element in every aspect of this. So it leaves you with the uneasy--a thought, feeling that the normal ways that we assess quality are deeply flawed, because they are a mix of luck and skill. So, where does that leave us? Guest: Well, the first thing is you need to realize the impact of this randomness. And you are correct--that there is so much that goes on that we attribute to skill that isn't skill. And there might be a sports example where somebody is on a so-called hot streak. And again, if you could be just purely random--that you've got a guy sitting there, throwing like 10 baskets in a row, or something like that. And it's really important to separate that out. So, how likely is it that something like this could happen? You'd be surprised at how likely it actually is. So, this is definitely the case. Yes it's true, even in scientific work: it is possible to make a discovery--like a real discovery--but it's random. You just get lucky in terms of what you actually do. In other fields of scientific research, to actually--and this is kind of interesting--the first person to publish, let's say a medical discovery, is often, they call it that there's a winner's curse. And basically what that means is that the first person to publish it, given all of the data mining that actually went on, it's likely that the effect is overstated, number one, or number two, doesn't exist. So when people replicate the study, they find that the effect is a lot smaller. Russ: try to-- Guest: Yeah. But this happens all the time, that it is so difficult to separate the skill from the luck. And I'm afraid that we don't really think it through. And we use these rules that are pretty naive.

Russ: So, Nassim Taleb has been a guest on EconTalk a number of times. And of course he's associated with issues related to these questions in his book Fooled by Randomness and in The Black Swan. I want to talk about black swans for a second. We're talking about: Is this a good strategy relative to other strategies? Another question we might ask as investors or as decision-makers is: What's the downside risk? I want to be prudent, I don't want to take an excessive risk. I don't mind sometimes if I make a little bit less. But what I don't want to have is, I don't want to be wiped out. I don't want to have catastrophic result. So, I worry, of course--I should worry also, not just about the average return but about that left-hand tail. Of a really catastrophic event. Can you talk a little bit about the role that assumptions play in assessing strategies, typically, in the financial literature? Taleb has argued that normality, the persistent presumption of normality, is a very destructive assumption because, although it makes things more tractable, it ignores these left-hand tail events when they are so-called "fat"--when they are more likely than they would be in a normal distribution. Can you talk about that for a little bit?

Guest: Sure. And this is actually part of my other research string, so it's very convenient that you mention this. So, we've been talking about something that is called the Sharpe Ratio. And that's basis the excess return divided by the volatility. So, it turns out that this is definitely not the metric that I would recommend using for evaluation of a strategy. And the reason is the following: That it does not take into account the tail behavior. So, it kind of--it assumes, directly, that the things that happen on the downside look approximately the same as the things that happen on the upside. So, it's symmetric. And often, you get this situation where you look across different investment styles, and you see that some investment high Sharpe Ratios, and other investment styles have very low Sharpe Ratios. And the reason is not that one investment is any better than the other. It's that there is a different sort of tail behavior. So, for example, the low Sharpe Ratio strategy might have the possibility, once in a while, of a big positive outcome. Like a lottery sort of payoff. Whereas the high Sharpe Ratio strategy, on average it does really well; but then there's a possibility of a catastrophic downside. So, the Sharpe Ratio is only useful if you are evaluating, as you said, relatively normally-distributed returns. Because it does not take into account the downside or upside differential. So that's something that I do not recommend. Russ: But it's a fascinating thing, because as an investor--and of course this comes up in social science as well--I don't live forever. I don't get an infinite number of draws from the urn; I don't get to play for a million years. So, I get the particular string that comes out from t_0 to t_1 in my time at the table. And of course that's a particular string. You can come back from Las Vegas and make a lot of money and think you are a great card player, when in fact you just happened to sit in on that string. And when you think about those asymmetric returns, I don't get the average return. I get whatever that happens to be in that time period when I'm holding that asset. I think one of Taleb's insights is that we think so often about normality, and so many things in life are normally distributed--height and weight and other things--that when we deal with these kind of problems we really don't have the apparatus, the intuition that we need to have. Guest: Yes. I totally agree. I think that Finance has done a particularly poor job. Indeed, if you look at the textbooks, the classic textbooks in Finance, there's very little mention of tail behavior. So, I'm always an advocate of holding a diversified portfolio, but that portfolio needs to be diversified over a number of dimensions. So, we usually think of getting that
portfolio with the highest possible projected return for some level of volatility. And my research points to: You need to take this third dimension into account, which is, the technical term for it is skewness. So, sometimes what's negatively skewed has got a downside tail that you don't like, and something that's positively skewed, like a lottery, has got a positive tail; and you need to take that into account when you form your portfolio. Because people have a preference. They prefer a strategy or a portfolio or an asset return that's got the positive tail. We want the big payoff. And we've got a distinct dislike for assets that have catastrophic downsides. So, that needs to be taken into account in designing and optimal portfolio strategy. And that is what I pushed in an article I published in the Journal of Finance a number of years ago. **Russ:** Yeah. If you put a dollar on red at the roulette wheel and you lose it, it's not a big deal. If you put your fortune on red--or a better way to say it is if you put a dollar on red and they can then go into your bank account, if red hasn't come up a certain number of times--the problem with the catastrophic thing is you can't bounce back. By definition you are in a hole that you either literally can't come back from--you are bankrupt--or, you are going to require a very, very, very long period of time to make your money back. **Guest:** Yeah. That's true. That's exactly how it works. Can I tell you a true story, something that happened to me recently? **Guest:** Sure. **Russ:** I get a phone call from a Duke graduate, who actually went to the Business School, and he wanted me to basically endorse a product that he'd been running for a few years. He was managing $400 million, a simple strategy, that he was buying S&P 500 futures, so he was kind of holding the market. But then, he was also adding on some options where he was writing or selling options that were out of the money for calls and puts. And when you do that, you collect a premium. And for the 5 years that he was operating, that premium led to an extra return. And it looked like he was beating the market. So every year he had about 2-300 basis points, or 2-3% above the S&P. And he basically said, Well, this is a great strategy; are you willing to endorse it? And I basically said to him--**Russ:** 'Are you out of your mind?' **Guest:** You didn't take my course. Because you would never ask me this question.' So, think about what that strategy is doing. So, that strategy, when the market goes up a lot, that means those options are valuable and you give up your upside. So, you have to pay off the person that you sold the option to. So you give up your upside in extreme up movements. And on the downside, if the market goes down, then you need to pay. So your downside is magnified if there is a big move in the market. So think about, so basically this person has changed the payoffs of the S&P 500, cut off the upside, and magnified the downside. And this extra return that he's getting over the 5 years, well, they are lucky over the 5 years. You haven't seen a big move up or down. So this is a great example of tilting the portfolio more toward this negative skewness. And when you've got negative skew, that means that the expected return should be higher. Because people don't like this downside possibility. So this kind of brings it together, that anything like this, whether it's an option or an insurance policy--insurance policy you pay a premium for and that's kind of a negative return for you, but if the fire happens and burns your house down, then you get the payoff. So you are willing to protect that downside. And I think that we don't really think this through enough in the way that we approach our portfolios.

38:33

**Russ:** Yeah, well, it's again, not just our portfolios. It's many, many things in life where we are evaluating the effectiveness of some strategy, and we don't like to think about it. It's a cherry-picking example: I want to think I'm doing the right thing and I look at how well it's going and then I can brag to my brother, hey, I made a killing. But I don't realize that the 5 years that I've been observing the data are not typical. And of course they never are. Almost by definition. There are things going on in any 5-year period that are distinctive, that you have to be careful about. So, to me, the lesson in this is you have to, when you are trying to evaluate quality of anything you have to have some intuition, what I would call wisdom, which is very hard, about the underlying logic of the strategy that the data itself don't speak. And I think this is a very dangerous problem in economics generally. So let me take it there and we'll look at some of the implications. So, Ed Learner's critique of econometrics and statistical significance is very similar to yours. Which is, if you run a lot of regressions, if you do a lot of statistical analysis and try all these various combinations of variables that you hope might show some significance, and then you find it, the classical measure of the t-statistic greater than 2 is meaningless. **Guest:** Yes. **Russ:** And the temptation to say, I found something' is so powerful, because you want to get published. And as you point out from your previous example, a lot of the findings aren't true. They don't hold up. **Guest:** So, it's not just getting published. My critique applies to people that are designing these quantitative strategies. I was—again, this is a true story. A number of years ago I was shown some research, at a high-level meeting, at one of the top 3 investment firms in the world. And this person was presenting the research, and basically he had found a variable that looked highly significant in beating the market, with a regression analysis, as you said. And it turned out that this variable was the 17th monthly lag in U.S. Industrial Production. **Russ:** Yeah. I've always known that's an important factor. But that's the beauty of his approach: nobody knows it, but by his fabulous deep look at the data, he uncovered this secret relationship that no one else knows. **Guest:** So, 17th lag? That seems a little unusual. So, usually we think of maybe the 2nd because one month the data isn't available because of the publication
delay. Maybe the 3rd. But the 17th--where's that coming from? And then he basically said, 'Well, that's the only one that worked.' So, it's the jelly bean example. There's a paper that's circulating that looks at the performance of stocks sorted by the first letter of the stock name. So, they'll look at the performance of all the companies that begin with the letter A, B, C, D. And one of them is significant. **Russ**: One of the 26. How shocking. **Guest**: Exactly. So this stuff happens all the time, whether it's in a very reputable investment bank or whether it's within academia. People basically are not adjusting for the data mining that is occurring.

42:36 **Russ**: Yeah. So, Learner's suggestion is to--he has more than one, but one of his early suggestions was to do some sensitivity analysis. Basically, look at all the combinations of the variables that you might look at and if you find that under most or all of them there's that you're kind of off effect size, or that you're trying to claim is the key variable--say, an example he gives in his paper, a beautiful example, is: Does capital punishment deter murder? And, does the threat of capital punishment induce people to stop killing other people? And he shows it's easy to find an analysis that shows that it does. And then of course it's easy to find one that shows that capital punishment increases murder--perhaps because, who knows why, more brutality in the culture, in our society. But it all depends on what you put on the right-hand side, what different variables you might include. But if after you did that, you found that it always deterred, or never did, then you'd feel more confident. So, you have a suggestion in Finance. What are your suggestions? **Guest**: Well, there are a number of suggestions that I explore in my research. One suggestion is to actually ditch the 2-sigma rule and move the cutoff higher just like and things like that. There are other approaches, too. The most popular approach is the so-called 'out of sample' approach, where you actually hold out some data; you fit your strategy to the past and then apply it to, let's say, the most recent 5 years of data to see if it actually works. And that is a long-established method. **Russ**: It sounds very good. **Guest**: It sounds good, but it's got problems. For example, you actually know what's happened in the past. So, if we hold out the last 5 years, well, we couldn't?I remember that we had this major global financial crisis? So the researcher knows that and might actually stick in some variables in the early part of the data that they know are going to work in the other sample. So, that's one problem. The other problem is a flawed scientific procedure where somebody looks at a model in sample and then takes it out of sample; it doesn't work. Then they go and basically re-do the model, removing some variables perhaps or a different method and then try another sample. It fails. And then they just keep on iterating back and forth, back and forth, until something works. And of course that is--you are just asking for the fluke situation. So, the final problem with the so-called out-of-sample technique is that you might fit a strategy over a number of years and then test it in the last 5 years; and the strategy might fail, but it might be a very good strategy and it's only because you have so few years, that just by bad luck the strategy fails. So, it is not a panacea to actually go to the other sample method.

46:16 **Russ**: So, some listeners will remember an interview that I did with Brian Nosek, the psychologist, a few years back, where he and others in psychology have become worried that some of the more iconic results in psychology are not replicable. And they've tried to replicate them. And some fail, some don't, obviously. But some seems to me a huge, enormous problem. As you say, it's one thing to talk about some particular psychology theory. When we are talking about people losing money for their personal retirement or when we talk, more importantly even about epidemiology where some claim about the relationship between, say, alcohol consumption and health, positive or negative, is going to maybe cost people's lives or save lives. The fact that many of these results don't stand up to replication seems to me to be an enormous problem in our scientific literature. It's a social science problem; it's a physical science problem. Do you agree that we've got a big problem there? That much of this so-called science is not scientific? **Guest**: I agree that there are problems but let me just elaborate a little bit. Psychology is at the very bottom of the hierarchy of science in terms of publishing results that are not significant. So, what I mean here is that it's rare in psychology to actually publish a result where you accept a hypothesis and the hypothesis garners no support. And you have a non-result. So, that's very difficult to publish in psychology. So the sort of papers that are published in psychology, over 92% of them are, 'Oh, here's a hypothesis; I did an experiment; and I get support.' So that is a problem. Because it leads to people essentially data mining to find a result to find a result and then getting it published. On top of that, you figure out that it's data mined if you replicate. And in psychology, there is a very poor culture of replication, people not that interested in replication of these experiments. And this contrasts, as you said, with medical science--epidemiology is a good example--where somebody actually might data mine the data and publish something, but then a half dozen other people replicate it and find that it isn't a fact. And we actually learn from that. And we actually learn something when that actually happens. In psychology, there isn't a large culture in terms of replication, but there's a particular reason for that. We're not running experiments with human subjects. We're actually looking at data, and the data is a fact. So, if you are looking at the New York Stock Exchange data, you are--everybody's got that data. So if you tell me that this particular value strategy has an excess return over the last 50 years of 5%, well, anybody can go in and immediately, with one line of computer code,
replicate that. So there isn't a large culture of replication because it really isn't necessary in terms of what we do. And psychology is a totally different game; and indeed they've had the data of, with people in the data of, these experiments-- Russ: Yeah, that's another-- Guest: and having to retract.

Russ: Yeah, that's a separate. But it seems to me that-- Guest: I mean it's a problem in Finance and in epidemiology. So, let me lay that case out and you can answer it. It's true everybody has got the New York Stock Exchange data. But that person who runs the 17th lagged Industrial Production variable and proves, using statistical techniques that, it's important, that, it's important to the model, well, is that going to work going forward? That to me is the replication, the equivalent of replication in that model. And similarly in macroeconomics--the cherry picking of sample size, of sample time period, of various variables to prove that Keynesianism works or doesn't work, that monetary policy is crucial or is irrelevant--to me it's just an intellectual cesspool. I hate to say it, but I don't see-- Guest: I totally don't agree here. Because my point is you need to adjust the significance for the number of tests. So, that person that ran the regression that the 17th lag of Industrial Production came in as significant--if they adjusted the significance level, given that they ran 24 different tests, they tried 2 years of lags, then that 17th variable is not significant. You would reject that variable. So my paper actually provides a method to avoid these mistakes. And again, this is a big deal. It can be somebody's pension money. Russ: Yeah. Guest: Running on a strategy like this. What I'm saying, you need to take into account that we've got, in this particular situation, 24 different things that have been tried. Not 1. And if you do that, then you minimize the chance that some bogus strategy based on a fluke finding is basically allocated to in your pension and you lose your money. So there are ways to deal with this, and my paper actually provides a method to do that.

Russ: Yeah, I understand that. But you happen to be sitting in a meeting that was an informal meeting and you were able to ask the question, 'How many times did you run that?' When I see--I'll give you just my favorite example. My favorite example was in epidemiology. There was a paper that showed that, front page of the New York Times; it was an enormous story--that drinking alcohol increased cancer among women. And that's a big finding. Without obviously you don't want to fool around with that. And unlike many of the journalists, I actually went and got the paper and I read it. And I contacted the researcher. And there were two things in the paper that just didn't get mentioned. One was the fact that they had the cancer history of the population in the sample; they had, 'Did your mother have cancer?' They had that information. But they did not use that in the analysis. I don't know why--since we know the genetic relationship. I don't know why they didn't use that. But more importantly was how they defined drinking and not drinking. They threw out all the people who didn't drink, on the grounds that people who say they don't drink in a survey might be used to drink. And then we'd be mismeasuring it. Well, that's true. It's also true that people who say they drink might have different drinking habits in the past. And of course once you throw out the non-drinkers, some of them--actually they had worse health. Not some of them--the average non-drinker had more cancer than the people who drank a little bit. So that was awkward. To throw those people seemed to be a rather unfortunate decision. They did it on the grounds that maybe those weren't measured accurately. Of course, others weren't measured accurately, either; they were all based on, say, memory, in this case, or whatever it was. It wasn't a lifetime sample. They were longitudinal observations. So, somebody publishes a paper in economics, and they don't tell you how many regressions they ran. Ever. Never. We don't get to see the video of what happened in the kitchen when they ran these tests and when they transfigured the variables and decided that the squared term was the right term. So, it seems to me that in the absence of that, we are really in, many of the things we find are not going to be replicated effectively. Guest: So, I 100% agree with you, with what you just said. That is a cesspool. That, what I was talking about earlier was fixing something relatively straightforward, where you know 24 tests have taken place. And the 17th lag works. So you can adjust for that: it's not significant. But what you are talking about is a broader critique that again, I mention in my research: That it's not just the number of tests. So, the other problem of that data. So, it might be that you start your analysis in 1971 versus 1970. That one year could make a huge difference in terms of your results. It might be that you trim outliers out of the data. It might be that you use an estimation method that has a higher chance of delivering a so-called significant result. So it's not just the number of tests, but it's all the choices that researchers make. And it is a very serious problem in academic research, because the editors of the scientific journals don't see all of the choices that have been made. It is also a problem in practical research in terms of the way that people are designing strategies for investors. However--and this is kind of, I think interesting, My paper has been very well received by investment bankers and people designing these strategies. And actually it's interesting because they actually don't want to market a strategy that turns out to be a fluke. Because that may hurt their reputation. It reduces the amount of fees that they get. And it really, basically it could reduce their bonus directly. So that actually has a strong incentive in terms of business practice to get it right. So, within the practitioner community, at least, there are strong incentives to reduce the impact of data mining, so that you can develop a good reputation. However, on the academic side, it's not as clear. As you said, there's minimal replication in some fields and the scientists don't see all of the hocus-pocus going on before the paper actually is submitted.
for scientific review. Russ: Yeah. When you were in that meeting at the investment bank and the person said it was significant and you said, 'Well, how many did you run?' and he said, 'Well, 26, 24,' whatever it was, and you said, 'That's not significant': Nobody around the table said, 'So what? Doesn't matter. We'll be able to sell it because it's over 2.' Guest: No, People, I'm sure: They do not want to do this. So that damages the reputation hugely. So, everything is reputation in terms of kind of street finance. And you want to do the right thing. You want to have in place a protocol--an explicit protocol--where some investor asked the question, 'Well, how do I know that this isn't due to data mining? And then what you can do is to point to your protocol, saying, 'We're well aware of data mining, And we actually take the following steps to minimize its impact. We obviously can't get rid of the chance that some findings could be a fluke, but we try our best to minimize that because we want to do the best thing for you, because that's how we make money.'

58:37 Russ: So, I was going to ask you--well we won't talk about it but I was going to raise it anyway. I was going to ask you whether you think Mike Krzyzewski, the coach at Duke University's basketball team is a good coach. He's considered one of the greatest coaches of all time. He's won over 1000 games. So, everyone knows he's a good coach. Or Bill Belichick of the New England Patriots. Everyone knows that person's a good coach. It seems to me--of course, there's a random element in those records, that deceives, that's complicated. And when we think about examples like finance or epidemiology, it seems like if we don't come down to the issue of what's really going on underneath the data, what's the model that you have in mind, that you are trying to measure--it seems like we're really lost. The reason I mention that is that, you know, in epidemiology, I don't really know the mechanism by which alcohol causes cancer. I hope some day we'll uncover it. But then just looking at statistical relationships without understanding the underlying biology seems to me to be dangerous. And similarly, if I don't understand why the 17th lagged measured level of Industrial Production is significant--it's not just that it's 26 tests and therefore not statistically significant. It's that: It doesn't make any sense. So, I think ultimately in all these cases when we are trying to assess the component of our outcomes that are due to randomness, we have to have some fundamental understanding of the causal mechanism or we are really at risk. Guest: I agree with that. So, I didn't need to adjust the significance level on the 17th lag of Industrial Production. That model is gone. It's history. And the employee that did it is probably history, also, after these comments. But it might be more complicated. It might be just that you see a strategy. I don't know really what's behind it. Because often--and then an investment banker or company might not be willing to reveal the inner workings of the model. Russ: Correct. Guest: So you need to have some sort of statistical method to actually do this. But I agree that the best thing you can do is to ask the question: 'What is the economic mechanism? Why does this work? Tell me the line of causality.' And try to minimize sort of spurious sort of relationships that are often put to the public where people are claiming causality when it's much more complicated. So I think that the bottom line here is that you do need to have a solid economic foundation. You need to have a story. Or you should be very suspicious of the performance of a particular strategy.

1:01:34 Russ: So, let's close with a thought on big data. There's an enormous, I think, seductiveness to data-based solutions, to all kinds of things, whether it's portfolio analysis, whether it's health. All kinds of things. We're much better at measuring things and we have a lot more measurements going on, and so a lot of data is being produced. And, there's a golden ring being held out there for us to grab that just, 'If we just use more data we're going to be able to improve our lives tremendously.' And of course, without data, you're in trouble. You don't want to just rely on your intuition. [more to come, 1:02:15]

Comments and Sharing

CATEGORIES: Campbell Harvey (1), Data and Evidence (21), Finance (60), Philosophy and Methodology (86)

TWITTER: Follow Russ Roberts @EconTalker

Like 6 6 8 + 1 0 1

Email this

POST A COMMENT Read comments

Name [Required]

Email Address [Required. Will not display.]

URL [Optional Begin with http://...]

Remember personal info? [Uses cookies]