Episode 2: Ron Wasserstein - Misunderstanding 'statistical significance'

Ron Wasserstein (RW): I use this example in my various talks to groups on P-Values. A small P-value is like a right swipe in Tinder. It means you have an interest. It doesn't mean you're ready to book a wedding venue.

Kaitlyn Regehr (KR): That was Ron Wasserstein, executive director of the American Statistical Association (ASA). An organisation whose job it is to promote the practice and profession of statistics.

Ron wasn't talking about the probability of meeting a potential spouse on a dating app, but rather the cautionary approach to the use of P-values, that we as researchers should apply.

RW: Uncertainty exists every place. It's just like the frigid weather in a Wisconsin winter. There are those that will flee from it, trying to hide in warmer havens elsewhere. Others will accept it and even delight in the omnipresent cold. These are the ones who buy the right gear and bravely take advantage of all the wonders of a challenging climate.

Significance tests and dichotomised P-values which we'll talk a lot about during this podcast, have turned many researchers into what I'll call, scientific snow birds, trying to avoid dealing with uncertainty by escaping to a 'happy place', where results are either statistically significant or they're not. In the real world, data provided a noisy signal. Variation, one of the causes of uncertainty, is that it's everywhere. Exact replication is difficult to achieve.

What we've argued, what I'll talk about during this podcast, is it's time to get the right statistical gear on and move towards a greater acceptance of uncertainty and embracing of variation.

[How Researchers Changed the World podcast introductory music]

KR: Welcome to How Researchers Changed the World, a podcast series supported by Taylor & Francis Group, which will demonstrate the real-world relevance, value and impact of academic research; and highlight the people and stories behind the research.

My name is Dr. Kaitlyn Regehr. I'm an academic researcher; an author and a scholar of digital and modern culture. I'm interested in how new technologies can broaden the reach and real-world impact of academic research.

In today's episode we're exploring The American Statistical Association's statement on P-values: Context, Process and Purpose.

P-values or probability values are a statistical evaluation tool used widely in research. The use of P-values in statistical hypothesis testing is common in many fields of research such as physics, economics, finance, political science, psychology, biology and sociology. Their misuse has been a matter of considerable controversy for decades. As Ron tells us, it's time to set the record straight.

RW: I'm Ron Wasserstein and I'm the executive director of The American Statistical Association, the worlds largest community of statisticians. We have members, statisticians, in over 90 counties. We provide membership services for statisticians. Statisticians are people who work in colleges and universities but also in government and in every manner of industry providing various kinds of really interesting statistical services

for people who need to understand their data better.

KR: Okay so what exactly is a P-value and what's the big problem?

RW: Why don't I start with a little bit about what a P-value is and why anybody should care? A P-value is, at the most basic level, a way of summarising the compatibility or incompatibility between a set of data has and a proposed model for that data. The model includes a whole bunch of assumptions.

One assumption that scientists like to call a Noel hypothesis and simply stated the smaller the P-value, the less compatible that data is with the model.

KM: Okay, that sounds a bit complex. Let's pull that apart a bit.

RW: That's pretty confusing isn't it. In fact, I would say that there's a decent chance that when I hear this podcast later, I'll discover I said something wrong in there. So, right there is one of the problems. P-Values are hard to explain properly and when you do, they aren't quite what you want to know. So, what happened is, over the years this P-value, it's a useful tool, let me say that to begin with. It's a very common and very useful tool for getting a handle on whether our data that we've collected fits or does not fit a model. Unfortunately, as you've gathered from my explanation, it's also a tool that is very easily misinterpreted.

It might be easy for someone to conclude that based on what I've said so far, that P-values are bad, they've always been bad. That's not true. P-values are a great tool. They were used very effectively for many decades to advance science in many fruitful ways. Unfortunately, they also began to be used inappropriately, to be misused and to be used less effectively.

KR: P-values and statistics go back a long way.

RW: P-value was really made popular around 1925 by the most famous statistician of the 20th century, R.A. Fisher.

KR: Sir Ronald Aylmer Fisher was a British statistician and geneticist. For his work in statistics, he has been credited as almost single-handedly created the foundations for modern statistical science.

RW: His idea was this. If as a result of your research, you got a small P-value then that was worth looking in to further. Unfortunately, he chose the word significant for that. An article in Scientific American a few years ago named the word, significant, as one of the seven most misunderstood words in science.

KR: Despite being a key weapon in the quantitative research arsenal, after years of controversy P-values were coming under widespread attack from various sources in the research community and in turn, have brought an entire field of statistics into disrepute.

RW: It really was right around 2010, for example we saw an article with the headline 'Odds Are It's Wrong, Science Fails to Face the Short Comings of Statistics'. That's a pretty galling headline. I went back this week and re-read that article and in it the author refers to statistics as a mutant form of mathematics. That's just painful.

KR: Yeah and statisticians don't like being referred to as mutants! It also had wider implications for the field at

large for example, in 2015 the Journal of Basic and Applied Social Psychology, banned the use of P-values and other forms of statistical methods as a result of these broad strokes misinterpretations.

RW: It was unfair and inappropriate and yet at the same time, our field had been writing for six decades about problems with P-values.

KR: For Ron and his colleagues at the ASA there was a lot at stake. Not only for the reputation of P-values for the future of statistics as a valued and trusted research methodology. The ASA, led by Ron, decided to embark on the largest consultation and evaluation of its kind.

RW: We were challenged to do the ASA statement on P-values because of these attacks on statistics as a whole field of research. Attacks on the misuse of P-levels were essentially being conflated with attacks on statistics as a field.

In April 2014, having been challenged by a colleague, my friend George Cobb challenged me to take up this issue. I went to the ASA board of directors and asked for some time and resources to take up this issue. Specifically, I asked the board for permission to start the process that might lead to the ASA writing a position paper on this issue.

It was not something that we had really done before, taking a position on an issue of statistical practice this broad. I felt like it was something we should consider doing and I asked the board if they agreed and I'm very grateful to the board for giving me room to start this process.

KR: The gauntlet had been laid down and the stage had been set; not for the vindication of P-values, but an open scientific discussion on how the ASA might provide guidance and a statement of their position on the matter.

RW: This is a unique situation and we felt after much deliberation that it took a special measure, because this is a specific aspect of statistics, statistical inference that goes well beyond ordinary statistical practice. It has been debated for so very long, we weren't going to go about this in a way that said, "alright, we're just taking a vote on science here and majority rules. If the 15 people voted this way in favour of P-values and 14 people vote this other way, then that's science." That's not science, it's never been science and God help us all if it becomes science.

What we decided to do was to get experts with a variety of opinions together and see what could be agreed upon. If there wasn't anything that could be agreed upon then so be it. If there were a set of principles upon which there was some agreement, then that's what we would put forward. Ultimately, and I'll say more about the process in a bit, but ultimately that's what was published in this article, some principles upon which this panel of experts felt that they could put their hand to, that they could put their names to and that we could be collectively comfortable with putting the ASA's name behind.

KR: Ron set out on a quest that would see hundreds of conversations, panels and meetings. Disagreements and reconciliations. He began the process of finding out what others had to say about the topic, and we asked him about this process.

RW: I was tasked by the board to identify people who were interested in a topic and to convene the group and try to see whether consensus could be brought together. It wasn't hard to identify people who were interested in the topic because lots of people have been writing on the topic for years.

I reached out to those people and I asked them about their interests, and I asked them to tell me who else was interested in the topic. I had no idea how people would react. I would say the most common reaction was, "I don't think there is a chance in hell that you'll get any kind of agreement on this, but I don't want to be

left out of the process, so I'm in."

KR: Ron is an organiser and quickly positioned himself as the 'ring master', and like any maestro conducting a massive orchestra or disparate voices, he set about forming teams, assigning people to those team. Setting timelines and setting up a clear and well-thought-out reporting mechanism. Ron was excited but did this meticulous approach work?

RW: I have to say that, that was the last point in time when that project was on the rails. After that for the longest time, nothing went right. This was way harder than I imagined it and for probably the next 16 to 17 months I wasn't sure whether there would be an ASA statement on P-values and statistical significance.

KR: This seems like a good point to pause our story and tell you about our supporting partner, internationally renowned publisher: Taylor & Francis Group. We've been working with Taylor & Francis to create a 12-week learning program to accompany this series; aimed at academic researchers looking to supercharge their career, it's entitled "How Your Research Can Change the World."

Working with thousands of academic researchers, Taylor & Francis have sought to make the journey of publishing as painless as possible, no matter what stage of your career.

This flexible programme will deliver the ultimate step-by-step guide to publishing your research, boosting your impact and building your profile.

Each week you'll receive a chapter via email. Over 12 weeks those chapters will build into an indispensable guide you can continue to use throughout your research career.

It's completely free and at the end of 122 weeks, you will receive a certificate and LinkedIn accreditation and have the opportunity to attend a Boot Camp organised by Taylor & Francis.

Interested? Head over to howresearchers.com

Before the break we heard Ron Wasserstein in uncertainty. In all probability his plan to unify clashing factions and publish a statement on P-values was in jeopardy. It was looking like the gamble would not pay off.

RW: This was an incredibly difficult topic on which to reach any kind of consensus. It was vastly more controversial that I realised it would be.

KR: Ron forged ahead. After months of challenges and against seemingly impossible odds, an end was in site.

RW: So, after many months of discussion, we gathered together right here in our offices in Alexandria, Virginia for two days of discussion. Not everybody involved in the process was able to join us, but we had a couple of dozen people here. Over the course of two days of discussion ably facilitated by my colleague, Regina Liu, we were able to hammer out six principles that we could agree on about P-values. That formed the frame work for the ASA statement that was ultimately published.

KR: Those six values are the following:

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a P-value passes a specific threshold.

4. Proper inference requires full reporting and transparency
5. A P-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a P-value does not provide a good measure of evidence regarding a model or hypothesis.

RW: We did that in November 2016. We spent another three months writing ultimately the article that became the ASA statement on P-values, hashing that out together. Even then I wasn't sure for another couple of months whether we would manage to get that agreed upon and signed off. We did, and now some three years later that article has been viewed over 300,000 times, which is huge by statistic standards. It's been cited over 1700 times which amounts to 11 citations per week over that three year period. That's extremely exciting. I feel like it's beginning to have an impact on science, which is what we really hoped for all along.

KR: The importance of this paper is hard to overstate. Underpinning many published scientific conclusions is the concept of "statistical significance," and this is typically assessed with a P-value index.

P-values are as widely used as they are misused. In an age of big-data, unprecedented scientific developments and an increase of complex datasets, properly conducted analyses and correct interpretation of statistical results also play a key role in ensuring that conclusions are sound and that uncertainty surrounding them is represented properly.

The impact of proper guidelines and the proper use and interpretation of P-values affect not only research, but research funding, journal practices, career advancement, scientific education, public policy, journalism, and law. So how has this impacted the research world?

RW: Our hope was, our hope still is and we're beginning to see that hope realised, that tool could be re-harnessed to be properly used and that other tools could be appropriately used. That the P-value could be reigned in and not have the oversized impact that it has. We're seeing that start to happen, through those citations people are recognising that the P-value needs to be properly used. With the new paper that we just released and a whole bunch of other papers that were released with it, we think that the major shift that we hope to see with regards to statistical inference, is taking place and that there will be a major change in how science uses statistics effectively in evaluating research.

KR: It would seem as though Ron's work has only just begun. The ASA has just published further guidance in the most recent edition of The American Statistician, which is open access and written for non-statisticians. The guidance is intended to go further and argues for an end to the concept of statistical significance and towards a model which the ASA have coined their ATOM Principle: Accept uncertainty, Thoughtful, Open and Modest.

The article spells out in great detail how those principles are intended to be applied. If you're planning on using P-values, you need to be looking into this and the ASA statement. You can find the link and summary to both on our website: howresearchers.com. There is also a full list of contributors.

The debate on this matter rages on and according to Ron, that's just the way it should be.

To find out more about this podcast and today's topic, visit howresearchers.com/pValues. Don't forget to subscribe to this podcast and follow us on Twitter, Facebook and LinkedIn @howresearchers

This podcast was written and produced by Monchü, recorded at Under the Apple Tree Studios. Our producers were Ryan Howe and Tabitha Whiting. Editing, mixing and mastering by Miles Myerscough-Harris at WBBC.

We would like to acknowledge the incredible support of Taylor & Francis Group with a special thank you to Elaine Devine and Clare Dodd.