

Tại sao các nhà thống kê lại trả lời những câu hỏi ngớ ngẩn mà mọi người chẳng bao giờ hỏi?

Matt Briggs*

“Mọi người không hiểu chúng tôi”, các nhà thống kê than phiền. Tuy nhiên, các nhà thống kê có tìm hiểu xem mọi người muốn gì không? Các nhà thống kê trả lời những câu hỏi ngớ ngẩn mà mọi người chẳng bao giờ đặt ra trong tâm trí họ, Matt Briggs nói. Thay vì, các nhà thống kê cần phải đưa ra những câu trả lời hữu ích cho những câu hỏi mà trên thực tế mọi người đặt ra.



Tôi biết rằng A đang thắng, nhưng tôi đặt cược cho B – một nhà thống kê nói với tôi

Phần lớn các thực hành thống kê được thiết kế để trả lời cho những vấn đề mà chẳng ai đặt ra và cũng chẳng ai muốn trả lời. Điều tồi tệ hơn là những người bình thường (những người không phải là các nhà thống kê) tin rằng những câu trả lời mà họ nhận được từ các phân tích thống kê là để trả lời cho những vấn

đề mà họ đã đưa ra. Tâm thế này vì thế đã làm họ tin tưởng vào những gì mà họ biết được hơn là những gì mà đáng ra họ phải thế.

Một nhà xã hội học muốn biết xem liệu có phải nhóm dân số A bị đối xử tệ hơn nhóm B, trong khi kiểm soát biến này hoặc biến kia. Một nhà khoa học chính trị quan tâm đến việc liệu chính sách A có giúp cho ứng cử viên của mình nhận được nhiều phiếu ủng hộ hơn so với chính sách B hay không. Hay một công ty muốn tìm hiểu xem loại thuốc nào là tốt hơn, A hay B. Vì vậy, họ thiết kế một nghiên cứu và thu thập dữ liệu, và kết quả dành cho công ty dược phẩm sẽ là quá nhiều bệnh nhân thấy tốt hơn khi sử dụng thuốc A, và cũng có quá nhiều bệnh nhân thấy tốt hơn khi sử dụng thuốc B. Lúc này, không một ai trên thế giới muốn biết khả năng là gì, với nhóm thử nghiệm này, A là tốt hơn B – tốt hơn với nghĩa là tỷ lệ bệnh nhân khỏi bệnh lớn hơn khi sử dụng thuốc A.

Chẳng ai muốn biết điều này bởi vì nó là một câu hỏi ngớ ngẩn, trả lời nó thật tầm thường. Chúng ta biết chính xác khả năng là gì, bởi vì sự kiện đã xảy ra rồi. Tất cả những gì chúng ta phải làm là nhìn vào các dữ liệu và xem hoặc là A tốt hơn B hoặc là B tốt hơn A. Xác suất mà A là tốt hơn, sự cố cho trước trong trường hợp này là hoặc bằng 1 hoặc bằng 0.

Câu hỏi đúng là ngớ ngẩn khi hỏi rằng có khả năng một con ngựa nào đó nhất định sẽ thắng trong cuộc đua ngựa năm 2011 ở Kentucky là bao nhiêu. Cuộc đua đã diễn ra trong tháng 5 vừa qua. Con

* Matthew Briggs là một nhà tư vấn thống kê và là giáo sư kiêm nhiệm về khoa học thống kê tại Đại học Cornell.

Animal Kingdom đã thắng trong cuộc thi đó. Bất kỳ người nào bây giờ tìm kiếm người nhận tiền cược để lấy lại tiền đặt cược của mình vào các con ngựa khác có thể phát điên.

Điều mà công ty thuốc muốn biết, trong điều kiện biết các bằng chứng thông qua các thử nghiệm và biết các bằng chứng khác về hai loại thuốc, là xác suất một tỷ lệ lớn các bệnh nhân trong tương lai sẽ có kết quả tốt hơn nếu họ sử dụng thuốc A thay vì sử dụng thuốc B là bao nhiêu? Họ có thể làm phức tạp câu hỏi này với những thông tin về các quy định, chi phí, sự cung cấp, hay thậm chí là quan điểm chính trị về các loại thuốc. Họ sẽ hỏi một vài điều chỉ có thể được quan sát thông qua các bệnh nhân mà họ vẫn chưa nhìn thấy.

Các thực hành thống kê cổ điển không trả lời những câu hỏi này, hay bất cứ câu hỏi nào giống như vậy. Thay vào đó, các câu hỏi gián tiếp (proxy question), những câu hỏi có nhiều điểm tương đồng nhất với câu hỏi thực tế, được hỏi. Vấn đề là ở chỗ những người bình thường không hiểu rằng sự tin tưởng vào việc trả lời những câu hỏi gián tiếp hiếm khi hoặc không bao giờ được chuyển thành sự tin tưởng về câu trả lời của các câu hỏi thực tế. Kết quả là thế giới đầy rẫy những tuyên bố hân hoan về sự bảo đảm đối với những phát hiện này hoặc phát hiện kia, tất cả cho rằng được ban phúc thông qua các số liệu thống kê - và tất cả những bảo đảm về sự vui mừng đó là phi lý.

Nếu một ai đó theo trường phái tần suất thực hành, câu hỏi gián tiếp là: xác suất có thể xảy ra là bao nhiêu, nếu cuộc thử nghiệm được thực hiện lại với số lần không giới hạn, và mỗi lần nó được thực hiện lại, một kiểm định thống kê được tính toán - một trong những kiểm định có thể được chọn - các thống kê kiểm định của chuỗi thử nghiệm lặp lại này có thể lớn hơn so với thống kê quan sát trong thực tế của những dữ liệu này, tất cả giả định là một vài mô hình xác suất là phù hợp và không có sai sót và giả định các tham số của mô

hình này được đặt bằng 0 (hoặc bằng một số nào đó). Như tôi đã nói lúc đầu, đây không phải là một câu hỏi mà ai cũng có thể hỏi trên thực tế, và không phải vì họ không thể kiểm soát được nó một cách liên tục.

Tất nhiên, câu trả lời cho câu hỏi này là giá trị thống kê kiểm định p (p -value), một khái niệm xác suất rất khác biệt với những trực giác mà không một người bình thường nào có thể có trong suy nghĩ của mình (cũng không phải phán đoán bằng cách tìm kiếm những điều đúng đắn trong sách giáo khoa là có thể hơn được một vài nhà thống kê). Chỉ cần hiểu rằng giá trị p -value nhỏ hơn nhiều so với một mức đã được thừa nhận, mức không cần phải hỏi, và vì thế kết quả là “có ý nghĩa thống kê”, người bình thường cho rằng câu hỏi ban đầu của anh ta đã được trả lời và khả năng A tốt hơn là cao.

Tuy nhiên, anh ta không nên tin điều này, bởi vì lẽ tất nhiên điều này có thể không đúng. Giá trị p -value có thể là nhỏ, nhỏ như mức bạn muốn, và xác suất A là tốt hơn có thể vẫn thấp. Điều này rất có thể xảy ra, dù rằng mô hình là phù hợp và kiểm định thống kê được chọn là “đúng” đối với vấn đề đang được nghiên cứu. Biết trước là không chắc mô hình là đúng, và biết trước là các kiểm định thống kê thường được lựa chọn để thuận tiện cho nhà thống kê chứ không phải cho người bình thường, vì vậy rất có thể là một người bình thường thậm chí còn có câu hỏi được trả lời sâu sắc hơn so với những gì anh ta nghĩ. Tuy nhiên, niềm tin của anh ta vẫn sẽ tiếp tục không suy giảm.

Tình trạng được cải thiện khi ta xem xét phân phối của các tham số của mô hình. Nhưng điều này không có nhiều. Câu hỏi gián tiếp trở thành: xác suất tham số đại diện cho sự cải thiện khi sử dụng A lớn hơn tham số đại diện cho B là bao nhiêu, biết trước là mô hình là đúng và biết trước các kết quả thử nghiệm? Câu hỏi này là phù hợp hơn với những gì mà một người bình thường mong muốn, nhưng nó vẫn là một sự lảng tránh. Sau cùng, những tham số không quan

sát được và không thể quan sát phải làm gì đối với thắc mắc của một người bình thường ?

Cũng như với giá trị p-value, một lần nữa gặp trường hợp là các thông tin về các giá trị của các tham số vẫn phóng đại bằng chứng về khả năng mà một loại thuốc là tốt hơn những loại thuốc còn lại. Mọi người có thể có hầu hết các thông tin chắc chắn về giá trị của một vài tham số hoặc của các tham số (bằng cách ước lượng hoặc kiểm định giả thuyết), nhưng điều này không có nghĩa là có một xác suất lớn rằng loại thuốc này là tốt hơn những loại còn lại.

P-value là một khái niệm rất khác biệt với những trực giác mà không một người bình thường nào có thể có trong suy nghĩ của mình. Có thể nhiều nhà thống kê cũng vậy

Vậy là, mặc dù sự khác biệt đằng sau các tham số là lớn, vẫn có thể có trường hợp là xác suất A tốt hơn là thấp. Bạn có thể dễ dàng tự mình nhận thấy điều này: giả sử rằng xác suất trong tương lai bệnh nhân sẽ khỏi bệnh có thể được mô hình hóa bằng phân phối nhị thức, với tham số cho A bằng 0,50001 và cho B bằng 0,5. Vậy là xác suất của tham số A là lớn hơn của B là 1, như vậy chắc chắn tham số A là lớn hơn tham số B, nhưng khả năng mà nhiều bệnh nhân hơn sẽ khỏi bệnh khi sử dụng A chỉ là vừa đủ trên 50% (không quan tâm tới số lượng các bệnh nhân mới được kỳ vọng).

Công bố xác suất các tham số là khác nhau hoặc khác với một vài con số nào đó chẳng hạn như số 0, và người bình thường nghe thấy những điều không cần thiết như: xác suất A tốt hơn cao như xác suất các tham số khác nhau. Tuy nhiên, những điều này không bao giờ như thế cả. Xác suất khác nhau của các tham số không bao giờ đồng nghĩa với xác suất khác nhau về khả năng có thể quan sát được. Trừ khi người bình thường được làm cho hiểu, sự tự tin thái quá của anh ta được bảo toàn. Và chúng ta không được quên rằng ta vẫn đang giả thiết mô hình là đúng.

Tình trạng được cải thiện một lần nữa, nhưng chưa được giải quyết, bằng việc chuyển sang thống kê dự báo. Theo ngôn ngữ của Bayes, đây là phương pháp biểu thị các phân phối dự đoán ở phía sau. Thuật ngữ này không tồn tại một cách tự nhiên trong ngôn ngữ tần suất, dù ở đó có thể có các phương pháp dự báo.

Các phương pháp dự báo loại bỏ các câu hỏi gián tiếp: họ nói trực tiếp với chúng tôi là khả năng A tốt hơn, đó chính là điều mà những người bình thường muốn biết, và họ nói với anh ấy theo ngôn ngữ phù hợp với nhận thức của anh ta. Nếu người bình thường muốn tìm hiểu xem khả năng các bệnh nhân tương lai sử dụng thuốc A tốt hơn so với sử dụng thuốc B tới hai, hoặc ba lần là bao nhiêu, với các điều kiện kiểm chứng là cố định, hoặc nếu anh ta muốn hỏi bất kỳ một câu hỏi nào về các kết quả có thể quan sát được của những bệnh nhân mới, thì các phương pháp dự báo sẽ cho anh ta biết, và cho anh ta biết một cách rõ ràng không phải theo cách khó hiểu.

Một lần nữa, và không thể tránh khỏi, ngay cả với một phương pháp dự báo, có sự không chắc về mô hình, và sự không chắc chắn của mô hình không bao giờ tự tìm ra được cách để đưa ra được kết quả đúng. Vậy là ngay cả với các số liệu thống kê dự báo, một người bình thường có thể mất đi sự quá tự tin của mình. Tuy nhiên anh ta sẽ bớt đi sự quá tự tin so với những gì anh ta thể hiện theo các phương pháp thông thường.

Vậy tại sao các phương pháp dự báo không được sử dụng thường xuyên? Phải, chúng vẫn được sử dụng. Chúng được sử dụng ở một số nơi và thường được sử dụng bởi những người không phải là các nhà thống kê. Các chương trình máy tính được sử dụng trong nhận dạng vân tay hay nhận dạng khuôn mặt là những ví dụ tốt của các mô hình dự báo, vì chúng là các mô hình được sử dụng trong kỹ thuật tự động hóa các loại, ví dụ như trong bộ lọc thư rác.

Một người bình thường muốn biết những bức thư điện tử nào có khả năng là thư rác - một câu hỏi dự đoán. Anh ta sẽ thấy thật là quá quắt khi xuất hiện một giá trị p-value trong một ngày thu thập các thư điện tử mà nội dung của chúng có thể là một giả thuyết không đại loại như “tất cả các thư điện tử của ngày hôm nay đều là thư rác” đã bị bác bỏ. Và anh ta cũng không muốn biết về độ chắc chắn của một vài tham số khó hiểu. Những gì mà anh ta muốn là một chỉ báo liệu từng thư điện tử một có phải là thư rác hay không. Nếu mô hình được đưa ra trong các điều kiện dự báo - ví dụ, dường như là thư điện tử đích thực bị phân loại nhầm là thư rác - thì người bình thường có thể đánh giá chính xác liệu mô hình có hữu ích đối với anh ta hay không.

Một lý do quan trọng tại sao các phương pháp dự báo không được sử dụng là chỉ một số ít người biết đến sự tồn tại của chúng. Chúng không được dạy trong các chương trình giảng dạy, cũng không được nói đến trong sách giáo khoa được sử dụng bởi phần lớn những người thực hành thống kê. Các phương pháp dự báo chỉ có thể tìm thấy trong rất ít các khóa đào tạo nâng cao, được đưa vào như những bài học tự chọn trong lý thuyết Bayes. Vì tất cả các khóa học chỉ bao gồm lý thuyết sâu về tần số, và chỉ mở rộng đến lý thuyết Bayes nếu có thời gian, nên phần thời gian dành cho những lý thuyết mới là rất ít.

Các bài học về các phương pháp dự báo cũng không phải là đơn giản. Các vấn đề liên quan đến toán học thường rất khó. Đây là một vấn đề vì các nhà thống kê lý thuyết khẳng định rằng các chủ đề của họ là một nhánh của toán học, và họ thiên về giảng dạy phương pháp mà không giảng giải về bản chất toán học đằng sau phương pháp đó. Ngoài ra một người bình thường cần rất ít hoặc không cần tới toán học để hiểu “khả năng A tốt hơn B là 80%” nghĩa là gì.

Ngoài ra, anh ta cũng không cần phải hiểu về môn giải tích để chạy các phần mềm thống kê - điều

mà không may cũng lại thiếu đối với các phương pháp dự báo. Không có chương trình phần mềm nào mà mọi người có thể mua hoặc tải về. Các nhà thống kê muốn sử dụng các công cụ dự báo buộc phải viết các chương trình riêng cho họ.

Nhiều nhà thống kê cần phải lưu ý hơn. Từ một vài ví dụ đã nêu, chúng ta thấy rằng tác dụng chủ yếu của các phương pháp dự báo là: tức thời và liên tục phản hồi về các kết quả của mô hình. Nếu như mô hình không tốt, chúng ta sẽ biết rất nhanh; sự thành thạo về thống kê sẽ không cần thiết để khẳng định điều này. Vì thế nó không dành cho phân tích thống kê các câu hỏi gián tiếp, nơi mà thường xuyên - đặc biệt trong các lĩnh vực như xã hội học, tâm lý học, chính trị và những lĩnh vực tương tự - tất cả những gì chúng ta có là các mệnh đề về p-value (khi nó có mức ý nghĩa cao), các mệnh đề mà không có bất kỳ gợi ý nào rằng mô hình có thể không hoàn hảo, và các mệnh đề cho độ chắc chắn sai thì các kết quả là chắc chắn như các tác giả nghĩ.

Tuy nhiên, các câu hỏi gián tiếp thường đưa ra các câu trả lời đúng, hoặc điều gì đó giống như vậy. Ví dụ, người ta không bị bất ngờ ngất đi (ít nhất cũng không phải là thường xuyên) vì những kiểm định và phân tích thuộc sử dụng các phương pháp dựa trên câu hỏi gián tiếp. Điều này là bởi vì các công ty kiểm định hai loại thuốc đã biết hoặc nghĩ rằng họ đã biết nhiều về các cơ hội tốt hơn của A so với B trước khi họ thực hiện thử nghiệm. Họ có nhiều hiểu biết về hóa học, sinh học, về sự phản ứng của những con người khác và có thể của một vài con chuột đã sử dụng thuốc, và nhiều hơn thế nữa. Thật vậy, họ biết rất nhiều về các loại thuốc trước khi cuộc thử nghiệm bắt đầu, cuộc thử nghiệm được thực hiện bởi vì tất cả những bằng chứng này là hoàn toàn thông nhất với ý kiến rằng A sẽ đánh bại B. Các câu hỏi gián tiếp đơn thuần chỉ để khẳng định những gì đã nghĩ ngờ.

Nhưng đó là các phương pháp dự báo mà hầu như kết hợp một cách tự nhiên nhất với các chứng cứ trước và cho phép chúng ta bù đắp cho những thiếu hụt của những hiểu biết trước đó. Họ làm việc này bằng cách định lượng trực tiếp những gì chúng ta có thể mong đợi trong tương lai; họ không dành thời gian để nói cho chúng ta biết câu chuyện về những gì đã diễn ra.

Những chỉ trích này về thực hành thống kê không chỉ giới hạn đối với thuốc hay các tiếp thị thử nghiệm, hay đối với việc nghiên cứu của các nhà xã hội học và các nhà khoa học chính trị. Những người trong các lĩnh vực này là những người có tiếng nói nhất trong số những người quá hiểu bản thân mình. Các phương pháp dự báo có thể được sử dụng cho bất kỳ mô hình thống kê nào, hay bất cứ thủ tục thông thường nào

được sử dụng. Chúng tạo thành một sự mở rộng cho các phân tích Bayes và áp dụng cho bất cứ lĩnh vực nào nơi mà chúng ta không biết chắc chắn về những điều mà chúng ta chưa được thấy.

Đến đây, hi vọng rằng nhiều phương pháp dự báo hơn sẽ được sử dụng, đặc biệt là trong những lĩnh vực được chính phủ quan tâm để đưa ra các quyết định về chính sách. Chính trong các lĩnh vực đặc biệt này, tôi hi vọng là tất cả chúng ta đồng ý rằng mọi người đều hiểu quá rõ về bản thân mình.

Đinh Bá Hiến (dịch)

Nguồn: "Why do statisticians answer silly questions that no one ever asks?", Significance Magazine, Volume 9 Issue 1 (February 2012), <http://www.significancemagazine.org/view/MTQ3NTI3MS9KQS84NjgzNDUvbnVsbA==/journalArticlePdf.html>