

TRONG NGÀNH KHOA HỌC THỐNG KÊ THẾ GIỚI **NHỮNG ĐÓNG GÓP QUAN TRỌNG CỦA NGƯỜI VIỆT**

Nguyễn Văn Tuấn ^(*)

LTS: Khoa học thống kê và những đóng góp của các nhà thống kê Việt Nam đang được độc giả quan tâm. Ban biên tập Tờ Thông tin khoa học thống kê sưu tầm qua mạng bài viết của TS. Nguyễn Văn Tuấn, giáo sư, tiến sĩ Viện Nghiên cứu y khoa Sydney, Australia để độc giả tham khảo.

Khoa học thống kê

Nhưng bộ môn thống kê mà tôi muốn bàn ở đây không phải là các hoạt động truyền thống như mô tả trên, mà là *khoa học thống kê* (statistical science), tức là một bộ môn khoa học thực nghiệm: phát triển giả thiết khoa học, tiến hành thí nghiệm, phân tích dữ liệu, và diễn dịch dữ liệu. Có người thường nghĩ rằng thống kê là một công cụ của khoa học, nhưng tôi nghĩ rằng quan điểm đó không chính xác, vì trong thực tế, nhà thống kê học không chỉ là người đơn thuần làm phân tích dữ liệu, mà là một nhà khoa học, một nhà suy nghĩ (“thinker”) về nghiên cứu khoa học.

Khoa học thống kê đóng một vai trò cực kỳ quan trọng, một vai trò không thể thiếu được trong bất cứ công trình nghiên cứu khoa học, nhất là khoa học thực nghiệm như y khoa, sinh học, nông nghiệp, hóa học, và ngay cả xã hội học. Thí nghiệm dựa vào các

phương pháp thống kê học có thể cung cấp cho khoa học những câu trả lời khách quan nhất cho những vấn đề khó khăn nhất.

Làm sao chúng ta biết phẫu thuật A có hiệu quả tốt hơn phẫu thuật B? Làm sao chúng ta biết aspirin có thể đem lại lợi ích cho bệnh nhân? Trong số 25 ngàn gen trong cơ thể con người, gen nào có khả năng gây ra ung thư, tiểu đường, loãng xương? Làm sao chúng ta biết một giống lúa mới có sản lượng cao hơn giống lúa cũ? Làm sao chúng ta biết được quá trình học vấn ở cấp phổ thông có ảnh hưởng đến kết quả học tập ở bậc đại học? Tại sao trẻ học sinh tiểu học ở nông thôn hay bỏ học? Làm sao chúng ta biết những đặc tính nào của cà phê được người tiêu thụ ưa chuộng, và có sự khác biệt về sở thích giữa nam và nữ hay không? Một số du khách vào Việt Nam có xu hướng “một đi không trở lại”, vậy yếu tố nào đã làm cho họ có xu hướng đó? Làm sao chúng ta biết người dân ủng hộ chính sách A mà không là chính sách B? Vân vân và vân vân. Đó là những vấn đề mà thống kê học có thể cung cấp câu trả lời khách quan và đáng tin cậy nhất.

Chẳng hạn như vấn đề bệnh teo cơ delta (có khi gọi là bệnh “chim sệ cánh”) mà ngành y tế nước ta đang đương đầu hiện

[*] TS Toán Thống kê, GSTS Y khoa Viện Nghiên cứu Y khoa Garvan, Sydney, Australia

nay. Cho đến nay dù đã xảy ra hơn 2000 trường hợp trên toàn quốc, mà các chuyên gia vẫn chưa biết chính xác nguyên nhân hay các yếu tố nguy cơ nào gây nên bệnh! Chính vì không có dữ liệu trong tay, cho nên có khá nhiều chuyên gia hàng đầu trong ngành đề xuất nhiều yếu tố nguy cơ dựa vào những phát biểu cá nhân và chung chung như “theo ý kiến của tôi”, hay “qua kinh nghiệm 50 năm hành nghề của tôi”, hay “tôi nghĩ rằng”. Nhưng trong khoa học, không có cái gọi là “theo ý kiến của tôi” hay “theo kinh nghiệm của tôi”, vì khoa học dựa vào dữ liệu thực tế được quan sát và đo lường chính xác (hay ít ra là khá chính xác) để phát hiện một yếu tố nguy cơ cho bệnh tật, hay nói chung là để phát biểu một định đề. Và, để có những dữ liệu đó, nhà khoa học phải tiến hành thí nghiệm.

Một thí nghiệm khoa học được bắt đầu bằng một ý tưởng, một giả thiết, và để thử nghiệm giả thiết đó, một qui trình khảo sát phải được tiến hành theo các bước chung như: thiết kế, thu thập dữ liệu, phân tích dữ liệu, và diễn dịch ý nghĩa của dữ liệu. Mỗi một bước trong qui trình đó đều có sự cống hiến quan trọng của thống kê học. Những câu hỏi then chốt đặt ra là: phải thiết kế một công trình thí nghiệm như thế nào, cần bao nhiêu bệnh nhân, có cần nhóm đối chứng (tức không bị bệnh) hay không, phương pháp thu thập dữ liệu như thế nào, phải đo lường cái gì, phân tích dữ liệu ra sao, v.v... Đó là “địa hạt” hoạt động của khoa học thống kê.

Những vấn đề trên chẳng những mang tính khoa học, mà còn mang tính đạo đức khoa học. Nếu không giải quyết thỏa đáng, có thể làm cho công trình nghiên cứu trở nên vô dụng và như thế nhà nghiên cứu có tội với bệnh nhân và tình nguyện viên.

Một trong những khía cạnh cực kỳ quan trọng trong nghiên cứu là xác định số lượng đối tượng hay bệnh nhân cần thiết để khảo sát. Một công trình nghiên cứu nếu không có đủ bệnh nhân hay tình nguyện viên, thì dữ liệu thu thập được sẽ không có giá trị khoa học cao. Ngược lại, nếu công trình nghiên cứu tuyển dụng quá nhiều bệnh nhân hơn số cần thiết, chẳng những gây ra phiền hà (có khi nguy hiểm) cho bệnh nhân một cách không cần thiết. Trong cả hai trường hợp, nhà nghiên cứu vi phạm đạo đức khoa học, và trong trường hợp nghiên cứu y khoa, đó là một vi phạm y đức. Do đó, hoạt động của khoa học thống kê không chỉ giới hạn trong các vấn đề khoa học, mà còn bảo vệ một khía cạnh của đạo đức khoa học.

Phân tích thống kê là một khâu quan trọng không thể thiếu được trong các công trình nghiên cứu khoa học, nhất là khoa học thực nghiệm. Một công trình nghiên cứu khoa học, cho dù có tốn kém và quan trọng cỡ nào, nếu không được phân tích đúng phương pháp sẽ không bao giờ có cơ hội được xuất hiện trong các tập san khoa học. Ngày nay, chỉ cần nhìn qua tất cả các tập san nghiên cứu khoa học trên thế giới, hầu như bất cứ bài báo y học nào cũng có phần “Statistical Analysis” (Phân tích thống kê), nơi mà tác giả phải mô tả cẩn thận phương pháp phân tích, tính toán như thế nào, và giải thích ngắn gọn tại sao sử dụng những phương pháp đó để hàm ý “bảo kê” hay tăng trọng lượng khoa học cho những phát biểu trong bài báo. Các tập san y học có uy tín càng cao yêu cầu về phân tích thống kê càng nặng. Không có phần phân tích thống kê, bài báo không thể xem là một “bài báo khoa học”. Không có phân tích thống kê, công trình nghiên cứu chưa được xem là hoàn tất.

Trong khoa học thống kê, có hai trường phái “cạnh tranh” song song với nhau, đó là trường phái tần số (frequentist school) và trường phái Bayes (Bayesian school). Phần lớn các phương pháp thống kê đang sử dụng ngày nay được phát triển từ trường phái tần số, nhưng hiện nay, trường phái Bayes đang trên đà “chinh phục” khoa học bằng một suy nghĩ “mới” về khoa học và suy luận khoa học. Phương pháp thống kê thuộc trường phái tần số thường đơn giản hơn các phương pháp thuộc trường phái Bayes. Có người từng ví von rằng những ai làm thống kê theo trường phái Bayes là những người thiên tài!

Để hiểu sự khác biệt cơ bản giữa hai trường phái này, có lẽ cần phải nói đôi qua vài dòng về triết lý khoa học thống kê bằng một ví dụ về nghiên cứu y khoa. Để biết hai thuật điều trị có hiệu quả giống nhau hay không, nhà nghiên cứu phải thu thập dữ liệu trong hai nhóm bệnh nhân (một nhóm được điều trị bằng phương pháp A, và một nhóm được điều trị bằng phương pháp B). Trường phái tần số đặt câu hỏi rằng “nếu hai thuật điều trị có hiệu quả như nhau, xác suất mà dữ liệu quan sát là bao nhiêu”, nhưng trường phái Bayes hỏi khác: “Với dữ liệu quan sát được, xác suất mà thuật điều trị A có hiệu quả cao hơn thuật điều trị B là bao nhiêu”. Tuy hai cách hỏi thoát đầu mới đọc qua thì chẳng có gì khác nhau, nhưng suy nghĩ kỹ chúng ta sẽ thấy đó là sự khác biệt mang tính triết lý khoa học và ý nghĩa của nó rất quan trọng. Đối với người bác sĩ (hay nhà khoa học nói chung), suy luận theo trường phái Bayes là rất tự nhiên, rất hợp với thực tế. Trong y khoa lâm sàng, người bác sĩ phải sử dụng kết quả xét nghiệm để phán đoán bệnh nhân mắc hay không mắc ung thư (cũng giống như trong nghiên cứu khoa học, chúng ta phải sử dụng số liệu để suy luận về khả năng của một giả thiết).

Cống hiến của người Việt

Trong những người Việt ở nước ngoài làm khoa học, số nhà khoa học thống kê cũng không nhiều, nếu không muốn nói là “đếm đầu ngón tay”. Và họ cũng là những nhà khoa học “trầm lặng”, bởi vì những công trình nghiên cứu của họ khó có thể trở thành một bản tin, một câu chuyện trên báo chí hay hệ thống truyền thông đại chúng. Phát triển một phương pháp xử lý số liệu, tuy có thể có ảnh hưởng cơ bản đến tất cả các ngành khoa học hàng trăm năm, nhưng khó mà hấp dẫn so với một khám phá về gen liên quan đến bệnh tật.

Tuy số nhà khoa học thống kê gốc Việt trên thế giới không nhiều, nhưng họ có nhiều đóng góp quan trọng (có khi rất quan trọng) cho khoa học thống kê. Trong số những người có những cống hiến quan trọng đó, phải kể đến giáo sư Phạm Gia Thụ tại Đại học Moncton ở Canada (mà *Người viết xứ* có một bài viết về ông - **"Đất nước đang thời kỳ xây dựng, đoàn kết, tiến lên"**) và giáo sư Huỳnh Huynh, thuộc Đại học South Carolina (Mỹ). Có lẽ nhiều người Việt Nam ở trong nước, kể cả giới khoa học, ít biết đến hai người này, nhưng trong giới thống kê học, họ là hai nhà khoa học có tiếng và có ảnh hưởng quan trọng. Ông Huỳnh Huynh thuộc trường phái tần số và ông Phạm Gia Thụ thuộc trường phái Bayes.

Như đề cập trong phần trên, trong các nghiên cứu khoa học, việc xác định số lượng đối tượng cần thiết cho công trình nghiên cứu cực kỳ quan trọng, vì nó chẳng những là vấn đề khoa học mà còn là vấn đề đạo đức khoa học. Đóng góp quan trọng của ông Phạm Gia Thụ trong lĩnh vực này xác định số lượng đối tượng nghiên cứu (hay xác định cỡ mẫu - sample size determination) theo lý thuyết của trường phái Bayes có thể

nói là cơ bản. Trong một bài báo đăng trên tập san *The Statistician* năm 1992 (1) giáo sư Thụ đã mở ra một hướng đi cho lĩnh vực này. Tiếp tục công trình năm 1992, ông Thụ còn cho xuất bản một số công trình có giá trị trên các tập san có uy tín cao trong khoa học thống kê như *Journal of the Royal Statistical Society* (Anh), *Statistics, Communications in Statistics-Theory and Methods, Mathematical and Computer Modelling*, v.v... Kể từ khi công trình nghiên cứu năm 1992, cho đến nay không một bài báo khoa học nào về xác định cỡ mẫu theo trường phái Bayes mà không nhắc đến bài báo của tác giả “Pham-Gia T” (tức ông Phạm Gia Thụ).

Trong khoa học, có một công trình nghiên cứu chẳng những được nhiều đồng nghiệp trên thế giới tham khảo, mà còn được đồng nghiệp tham gia bình luận, khen ngợi là một vinh dự lớn. Công trình của giáo sư Thụ là một công trình như thế: được tham khảo rất nhiều lần, và được ban biên tập tập san mời đồng nghiệp trên thế giới tham gia bình luận. Cho đến nay, nếu một nhà khoa học nào mới bước vào nghiên cứu về lĩnh vực này mà “quên” không nhắc đến tác giả Pham-Gia thì chắc chắn sẽ bị người bình duyệt nhắc nhở ngay: đề nghị tác giả tham khảo công trình của Pham-Gia! Khi một công trình khoa học về lĩnh vực xác định cỡ mẫu theo trường phái Bayes, ông còn được ban biên tập tập san mời bình luận (2).

Trong nhiều nghiên cứu khoa học, nhà nghiên cứu phải so sánh nhiều nhóm đối tượng, và mỗi nhóm thường được thử nghiệm (đo lường) nhiều lần. Chẳng hạn như trong nghiên cứu về hiệu quả của hai loại thuốc trong việc điều trị loãng xương, nhà nghiên cứu phải có hai nhóm bệnh nhân, mỗi bệnh nhân sẽ được mời tái khám nhiều lần để nhà nghiên cứu thu thập dữ

liệu. Trong thuật ngữ thống kê, người ta gọi đó là nghiên cứu dạng *repeated measure design* (thiết kế tái đo lường). Một trong những khó khăn trong việc phân tích các dữ liệu thu thập từ những nghiên cứu này là có nhiều giá trị cho mỗi bệnh nhân. Vào đầu thế kỷ 20, giáo sư Ronald Fisher, nhà thống kê học người Anh và cũng là “cha đẻ” của khoa học thống kê ứng dụng, có đề xuất một phương pháp phân tích có tên là *phân tích phương sai* (analysis of variance) cho các thí nghiệm có nhiều nhóm đối tượng, mà sau này sách giáo khoa gọi là *kiểm định F* (lấy chữ cái của tên ông Fisher đặt tên cho phương pháp). Nhưng phương pháp của ông Fisher có vấn đề khi ứng dụng vào nghiên cứu tái đo lường vì sự tương quan giữa các giá trị đo lường trong mỗi đối tượng nghiên cứu không được điều chỉnh thỏa đáng. Vấn đề này kéo dài mãi đến năm 1970, đến khi có công trình của ông Huynh và đồng tác giả Feldt. Trong một bài báo chỉ 7 trang rất quan trọng đăng trên tập san số 1 của khoa học thống kê, *Journal of the American Statistical Association* hay JASA (3), hai ông Huynh và Feldt đã giải quyết vấn đề này bằng cách chỉ ra điều kiện cần thiết để kiểm định F có ý nghĩa thống kê.

Công trình của ông Huynh và Feldt gây một tiếng vang cực kỳ lớn trong khoa học thống kê, đến nỗi bất cứ phần mềm máy tính nào, bất cứ sách giáo khoa nào, bất cứ bài báo nào về phân tích phương sai tái đo lường mà không tham khảo hay nhắc đến bài báo năm 1970 của hai ông. Rất ít công trình nghiên cứu nào trong thống kê học lại có ảnh hưởng sâu rộng như thế, và cái tên Huynh và Feldt nay đã trở thành những cái tên quen thuộc: “The Huynh-Feldt epsilon” hay “Two-factorial Huynh-Feldt test”. Bạn đọc có thể gõ “Huynh-Feldt” trên Google sẽ thấy bao nhiêu bài báo nói về phương pháp này!

Đến năm 1976, hai ông Huynh và Feldt lại công bố một công trình nghiên cứu chỉ 13 trang gây thêm tiếng vang trong ngành mà sau này người ta hay nhắc đến với thuật ngữ “The Huynh-Feldt correction” (4). Ngoài hai công trình tiêu biểu này, ông Huynh còn có nhiều đóng góp trong lĩnh vực thẩm định giáo dục bằng phân tích thống kê được công bố trên các tập san số 1 trong ngành tâm lý học và thống kê tâm lý học như *Psychometrika*, *Psychological Bulletin*, hay nghiên cứu thống kê giáo dục (*Journal of Educational Statistics*, *Journal of Educational Measurement*).

Khoa học thống kê ở Việt Nam

Trong vòng trên dưới 100 năm qua, thống kê học đã nhanh chóng tiến vào tất cả các lĩnh vực nghiên cứu khoa học, và trong quá trình chinh phục, thống kê học tạo nên những bộ môn nghiên cứu mới. Các bộ môn đó có thể kể đến như *biometry* (sinh trắc học), *technometrics* (kỹ thuật trắc học), *bioinformatics* (thông tin học), *psychometry* (tâm lý trắc học), *anthropometry* (nhân trắc học), v.v... Thật vậy, khoa học thống kê đã chi phối đến tất cả các bộ môn khoa học, và ảnh hưởng của các nhà thống kê đã lan tràn đến mọi bộ môn khoa học với sự chinh phục nhanh chóng sánh ngang hàng với Attila, Mohammed, và loài bọ Colorado” (“Statisticians have already over-run every branch of science with a rapidity of conquest rivalled by Attila, Mohammed, and the Colorado beetle” (Maurice Kendall, 1942).

Tuy khoa học thống kê đã góp phần tạo nên diện mạo khoa học hiện đại ngày nay, nhưng ở nước ta, khoa học thống kê còn rất kém. Trong khi bất cứ đại học nào ở các nước tiên tiến đều có một bộ môn thống kê học, một bộ môn chuyên cung cấp tư vấn về thống kê học cho các nhà khoa học thực nghiệm, thì ở nước ta, các bộ môn khoa học

vẫn còn giới hạn ở khoa toán, và chỉ xoay quanh một vài vấn đề căn bản thống kê.

Sự “lạc hậu” về khoa học thống kê ở nước ta đã làm ảnh hưởng không nhỏ đến khoa học nước nhà. Vì các bộ môn khoa học thực nghiệm ở nước ta chưa được sự hỗ trợ từ khoa học thống kê, cho nên rất nhiều nghiên cứu khoa học ở nước ta chưa có chất lượng cao, và chưa thể công bố trên các tập san khoa học quốc tế. Người viết bài này biết rất nhiều trường hợp nhiều nghiên cứu từ Việt Nam chỉ vì không được thiết kế đúng phương pháp, hay việc phân tích dữ liệu chưa đúng tiêu chuẩn khoa học nên bị các tập san khoa học từ chối công bố. Hệ quả là sự hiện diện của khoa học Việt Nam trên trường quốc tế còn quá khiêm tốn. Chỉ tính trong ngành y sinh học mà thôi, trong vòng 40 năm qua, số lượng bài báo từ các nhà khoa học ở Việt Nam chỉ trên dưới con số 300. Con số này cực kỳ khiêm tốn nếu so với 5.000 từ Thái Lan hay trên 20.000 từ Singapore.

Như đã đề cập phần trên, một công trình nghiên cứu dù có tốn bao nhiêu tiền, bao nhiêu năm tháng, mà không được thiết kế đúng và không được phân tích đúng phương pháp thì không thể xem là “khoa học”. Muốn thiết kế nghiên cứu đúng và phân tích dữ liệu đúng tiêu chuẩn, cần phải có sự đóng góp của khoa học thống kê và nhà thống kê học. Nhưng Việt Nam thiếu các nhà khoa học thống kê. Tuy nước ta có một viện toán học bề thế và có uy tín trên trường quốc tế, thì những môn học thiết thực nhất và căn bản nhất về thống kê ứng dụng lại nhờ các chuyên gia nước ngoài vào giảng dạy ngắn hạn. Mấy năm gần đây, trong quá trình hội nhập với thế giới, chúng ta phát hiện rằng khoa học nước ta còn kém về thống kê ứng dụng, và các cơ quan tài trợ học bổng phải gửi sinh viên ta đi học ở nước ngoài, kể cả ... Thái Lan!

Do đó, để nâng cao chất lượng nghiên cứu khoa học ở nước ta, một chiến lược quan trọng cần đặt ra là phát triển khoa học thống kê trong các trường đại học. Chúng ta cần rất nhiều nhà khoa học thống kê trong bất cứ lĩnh vực nghiên cứu nào. Ở nước ta, để phát triển khoa học nói chung và công nghệ sinh học nói riêng, chúng ta cần một đội ngũ nhà khoa học thống kê hơn là cần một đội ngũ nhà toán học. Người viết bài này tin rằng cùng với thế mạnh của toán học Việt Nam hiện nay, đội ngũ các nhà khoa học thống kê gốc Việt ở nước ngoài có thể và có khả năng đóng góp một phần công sức cho chiến lược phát triển khoa học nước nhà■

Chú thích

(1) T. Pham-Gia and Turkkan. Sample Size Determination in Bayesian Analysis. *The*

Statistician 1992; 41(4), 389-397 - with: Comments by C. Adcock, 399-404.

(2) Pham-Gia T. Sample Size Determination in Bayesian Analysis: a Commentary. *The Statistician*, JRSS, Series D 1995;44:163-166.

(3) Huynh H, Feldt LS. Conditions under which mean square ratios in repeated measurement designs have exact F distributions. *Journal of the American Statistical Association* 1970; 65:1582-1589.

(4) Huynh H, Feldt LS. Estimation Box correction for degrees of freedom from sample data in the randomized block and split-plot designs. *Journal of Educational Statistics* 1976; 1:69-82s.

Xin nói thêm rằng người viết bài này chỉ ngưỡng mộ công trình nghiên cứu, nhưng chưa quen biết và chưa bao giờ gặp mặt hai nhà khoa học thống kê đề cập trong bài viết.

HMT (Sưu tầm)