TƯƠNG LAI CỦA THỐNG KÊ HỌC

(Tiếp theo)

3. THỐNG KÊ NÒNG CỐT

Ngoài lĩnh vực hợp tác thì hoạt động chính của các nhà Thống kê là xây dựng các công cụ tính toán toán học hay các khái niệm được sử dụng để khai thác thông tin. Nhiều nghiên cứu dựa trên lý thuyết xác suất toán học cơ bản , tuy nhiên mục tiêu cuối cùng vẫn là để cung cấp các kết quả hữu ích cho công tác thực nghiệm. Đây chính là sự khác biệt giữa các kết quả nghiên cứu lý thuyết của các nhà Thống kê với phần lớn các lĩnh vực Toán học vốn theo đuổi các kết quả nghiên cứu không thực tế chỉ để tìm ra ý nghĩa vốn có của các kết quả này. Như đã nêu trong Báo cáo Odom:

Thống kê luôn gắn liền với các ứng dụng, và thậm chí trong Thống kê lý thuyết thì ý nghĩa của kết quả Thống kê cũng phụ thuộc nhiều vào việc các kết quả này liên quan tới loại hình ứng dụng nào. Điều này khác xa với tất cả các ngành Khoa học Toán học khác, ngoại trừ ngành Toán tin.

Thuật ngữ "Thống kê nòng cốt" không được các nhà Thống kê sử dụng thường xuyên nhưng lại rất hữu ích để mô tả chính xác ý nghĩa mà ta muốn nói tới. Ta xác định Thống kê nòng cốt là một bộ phận của hoạt động Thống kê và bộ phận này tập trung chủ yếu vào Thống kê chứ không phải nhằm vào nhu cầu về số liệu Thống kê của các lĩnh vực khoa học cụ thể. Đồng nghĩa với "nòng cốt" là "bên trong" cũng có thể dùng được. Điều này cho thấy Thống kê nòng cốt trái ngược với Thống kê bên ngoài. Hiểu theo nghĩa này thì các nhà Thống kê hoạt động trong cả Thống kê bên trong và Thống kê bên ngoài.

Các nghiên cứu về Thống kê nòng cốt tập trung vào việc phát triển các mô hình, phương pháp và lý thuyết Thống kê dựa trên các nguyên tắc chung của lĩnh vực này. Mục tiêu của Thống kê nòng cốt là nhằm tạo ra các triết lý, khái niệm, các phương pháp Thống kê và các công cụ tính toán thống nhất. Mặc dù đây là hoạt động bên trong như đã đề cập trên đây, tuy nhiên triết lý trung tâm của Thống kê nòng cốt chính là: một vấn đề quan trọng không phải vì nó hay nó đẹp (như trong toán học trừu tượng) mà là vì những tiềm năng ứng dụng rộng rãi của vấn đề đó hoặc có thể là ý nghĩa của nó trong việc tìm hiểu về hiệu quả của các phương pháp mà chúng ta sử dụng.

Thông qua việc xem xét cả các hoạt động bên trong và bên ngoài như trên ta có thể thấy rằng Thống kê nòng cốt phục vụ đắc lực như là một trung tâm thông tin. Thống kê nòng cốt được xác định bởi khả năng liên kết với hầu hết các ngành khoa học khác cũng như được sử dụng đồng thời trong các ngành đó. Việc những khái niệm và hệ phương pháp của Thống kê nòng cốt có thể được sử dụng đồng thời trong hầu hết các ngành khoa học và ứng dụng là yếu tố chủ yếu mang lại hiệu quả cho ngành Thống kê, từ đó mang lại nhiều giá trị cho tất cả các ngành khoa học.

Nghiên cứu Thống kê nòng cốt có thể trái ngược với các “Nghiên cứu Thống kê ứng dụng chuyên sâu” - là những nghiên cứu được điều chỉnh chặt chẽ hơn bởi nhu cầu phân tích số liệu để giải đáp những nghi ngờ trong một lĩnh vực khoa học cụ thể. Rõ ràng, loại nghiên cứu ứng dụng này dựa trên những kiến thức trọng tâm về các công cụ cũng như nắm bắt được

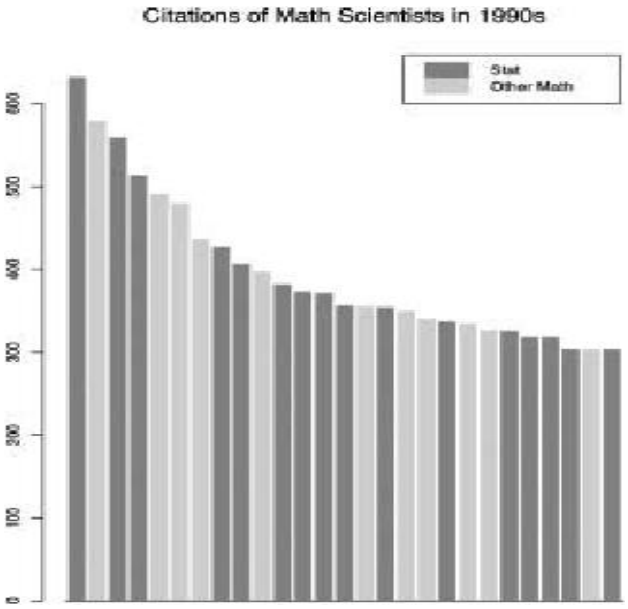
những hạn chế của các công cụ đó. Nó cũng cung cấp các tài liệu chưa qua xử lý cho nghiên cứu nòng cốt trong tương lai thông qua những nhu cầu chưa được đáp ứng.

3.1 Tìm hiểu về tính tương tác trong Thống kê nòng cốt

Theo bài phát biểu của Tiến sĩ Johnston tại Hội thảo, để nhận thấy được các hoạt động Thống kê nòng cốt mang lại giá trị lớn cho cộng đồng khoa học, có thể xem xét các số liệu được trích dẫn trong các tài liệu Thống kê. Tiến sĩ Johnstone cũng cảnh báo rằng không nên quá phụ thuộc vào số liệu trích dẫn bởi nhiều trích dẫn cho các bài viết cá nhân không phản ánh được chất lượng hay tầm quan trọng vốn có của số liệu. Mặc dù vậy, việc trích dẫn số liệu vẫn được đề xuất xét đến bởi nó đưa ra một thước đo đơn giản và dễ tiếp cận về tầm ảnh hưởng rộng rãi của nghiên cứu Thống kê trên các lĩnh vực khoa học khác.

Viện Thông tin Khoa học (ISI)- nơi đưa ra Danh mục Trích dẫn Khoa học và các tài liệu liên quan đã lên một số danh sách các nhà khoa học được trích dẫn nhiều nhất trong những năm 1990. Dựa trên các số liệu được ông Jennifer Minnick thuộc ISI cung cấp (ngày 11/10/2000), 18 trong số 25 nhà khoa học thuộc các ngành Toán học được trích dẫn nhiều nhất trong giai đoạn 1991-2001 là các nhà Thống kê hoặc Thống kê Sinh học. Hình 2 đưa ra số lượng trích dẫn trên mỗi tác giả. Bên cạnh đó, Tạp chí chuyên đề của Hiệp hội Thống kê Mỹ chắc chắn là tạp chí Khoa học Toán học được trích dẫn nhiều nhất. (Ghi chú: Theo báo cáo đầy đủ thì số liệu về các trích dẫn còn cho thấy một số lượng lớn hơn các nhà Thống kê được trích dẫn nhiều nhất, truy cập địa chỉ Website: http://in-cites.com/top/2003/index.html để xem danh sách 10 nhà nghiên cứu được trích dẫn nhiều nhất, trong đó tất cả đều là các nhà Thống kê tại thời điểm lên danh sách).

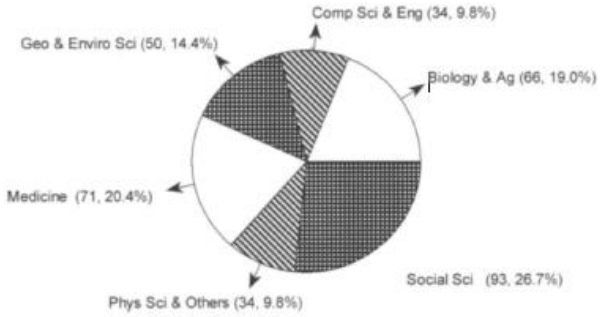
Hình 2: Số lượng trích dẫn của các nhà Toán học được trích dẫn nhiều nhất



Rõ ràng là việc trích dẫn nhiều bài viết Thống kê liên quan đến ngành Toán học nói chung phần nào cho thấy tầm ảnh hưởng mạnh mẽ của ngành Thống kê đối với các ngành khoa học. Ví dụ như bài viết của hai tác giả Hall và Titterington (năm 1987) nhằm xem xét vấn đề nan giải trong việc lựa chọn một thông số chung để đánh giá hàm không có tham biến. Trong bài báo này, khoảng 2/3 các trích dẫn không thuộc Thống kê nòng cốt là từ các tạp chí của Viện Kỹ thuật Điện và Điện tử, Tạp chí Kỹ thuật Hiển vi (Journal of Microscopy), Tạp chí Kỹ thuật Y sinh (Journal of Biomedical Engineering) và Tạp chí Vật lý học (Journal de Physique).

Một trong những bài viết quan trọng nhất hoàn toàn vượt ra khỏi nghiên cứu Thống kê nòng cốt và tập trung vào xu thế của các các lĩnh vực khoa học là bài viết của nhà Thống kê học Bradley Efron (năm 1979) giới thiệu về phương pháp Bootstrap. Theo dõi 500 trích dẫn trong bài viết này cho thấy chỉ có 152 trong số các trích dẫn đó nằm trong các tài liệu Thống kê. Hình 3 chỉ ra sự phân bố của các trích dẫn trong bài viết của Bradley Efron về phương pháp Bootstrap.

Hình 3: Sự phân bố giữa các trích dẫn về phương pháp Bootstrap của Bradley Efron thuộc các ngành khác nhau



Trong 500 trích dẫn trong bài viết của Bradley Efron, có 152 trích dẫn thuộc các tài liệu Thống kê

Tất nhiên, Thống kê nòng cốt cũng tạo ra các phương pháp có ý nghĩa lớn và rất hữu ích cho khoa học bởi nó hướng tới các ngành cụ thể, tìm ra các ý tưởng quan trọng và xây dựng những khái quát cần thiết để mở rộng ứng dụng. Ví dụ như xét về sự phát triển của các phương pháp tính tiền bảo hiểm dựa trên Tỉ lệ tử vong của các nhóm độ tuổi nhất định.

Trong năm 1972 và 1975, các ý tưởng về hồi quy các rủi ro tương xứng và phân tích khả năng từng phần đã được giới thiệu, làm phong phú thêm các công cụ sẵn có cho việc phân tích các số liệu theo thời gian khi có các số liệu đã được kiểm duyệt và thông tin đồng tham số. Từ đó, các ý tưởng và phương pháp này đã phát triển và lan rộng qua các ngành khoa học đến tất cả các nơi số liệu được kiểm duyệt hoặc được quan sát từng phần. Trong đó, lấy ngành Thiên văn học làm ví dụ: trong ngành Thiên văn học, ta có thể nhìn thấy một ngôi sao với một dụng cụ đo đạc nhưng không thể nhìn thấy ngôi sao đó với một dụng cụ đo đạc khác vì dấu hiệu không tương xứng.

Ví dụ cụ thể về sự tương tác

Ví dụ sau đây sẽ giải thích rõ hơn về việc các nghiên cứu Thống kê nòng cốt sử dụng và tương tác với các hoạt động Thống kê liên ngành ra sao. Ít nhất là từ một số hoạt động do NSF tài trợ, có thể phần nào thấy được loại hình tương tác nào nên duy trì nhằm hỗ trợ cho nghiên cứu Thống kê nòng cốt.

Năm 2001, ba nhà Vật lý thiên văn là Miller, Nichol và Batuski trong Tạp chí Science đã thừa nhận lý thuyết Big Bang về sự hình thành vũ trũ. Họ đã nghiên cứu về dấu hiệu gọi là dao động âm thanh trong phân bố vật chất của vũ trụ ngày nay, qua đó cho thấy dấu hiệu này hoàn toàn phù hợp với sự phân bố của Bức xạ phông vi sóng vũ trụ trước đây. Điều này không chỉ củng cố cho lý thuyết Big Bang mà còn cung cấp những kiến thức vật lý học về vũ trụ trước kia, qua đó có thể dự đoán được sự phân bố vật chất của vũ trụ trước đây cũng như trong tương lai dựa vào Bức xạ phông vi sóng.

Phát hiện này có được là nhờ sử dụng một phương pháp Thống kê mới gọi là Tỷ lệ phát hiện sai (FDR) để phát hiện các dao động. Với Tỷ lệ phát hiện sai là 1/4 thì 8 dao động bị đánh dấu là có khả năng không phù hợp với một phổ năng lượng phẳng không đặc trưng. Phân tích thêm nữa các tác giả đã đi tới kết luận rằng: các dao động là điểm khởi đầu quan trọng để phục vụ cho thống kê từ một phổ năng lượng có mật độ vật chất không đặc trưng.

Phương pháp này được phát triển với sự hợp tác của hai nhà Thống kê và được xuất bản trong Tạp chí Thiên văn học (Miller và cộng sự, xuất bản năm 2001). Với phương pháp này, các tác giả Miller, Nichol và Batuski có thể phát hiện và công bố phát hiện của mình trên Tạp chí Science (năm 2001) trong khi các nhóm nghiên cứu khác vẫn còn phải xoay xở với rất nhiều số liệu.

Thành công này cũng rất đáng quan tâm bởi nó cho ta thấy "Trung tâm thông tin" (Thống kê nòng cốt) hoạt động ra sao. Hình 4 minh họa trình tự của ý tưởng Thống kê.

Hình 4: Minh họa trình tự của ý tưởng Thống kê được khái quát trong Thống kê nòng cốt và áp dụng vào các lĩnh vực mới

Khi kiểm định các giả thuyết của cùng một bộ số liệu thì phải điều chỉnh các mức ý nghĩa của các kiểm định này để tránh khỏi việc loại bỏ sai các giả thuyết không. Vấn đề "suy luận đồng thời" này có lẽ là vấn đề được quan tâm nhất trong Thống kê Y tế, hoặc ít nhất là trong tất cả các tài liệu tham khảo được trích dẫn trong các tài liệu về Y tế. Thật vậy, đóng góp chủ yếu về Thống kê ở đây không phải là đưa ra phương pháp P-value liên tục cho mỗi sai số chuẩn đã được sử dụng trong ví dụ này. Phương pháp này thực sự đã truy nguyên đến phương pháp Simes trong những năm 1980 (hoặc có thể sớm hơn). Đóng góp chủ yếu mà ta muốn nói đến chính là đã tạo ra một giải thích mang tính lý thuyết đầy thuyết phục, chính là kiểm soát tỉ lệ phát hiện sai, từ đó các nhà nghiên cứu khác có thể đưa ra một dạng ước lượng.

Việc đưa ra ước lượng đã nhận được sự quan tâm của các nhà nghiên cứu bởi nó tạo ra khả năng lựa chọn ngưỡng trong các phương pháp khử nhiễu (Wavelet shrinkage) để xử lý tín hiệu Thống kê. Các nhà Thống kê học tại trường Đại học Carnegie Mellon (CMU) đã bắt đầu nghiên cứu về Tỷ lệ phát hiện sai FDR- vừa là vấn đề của Thống kê nòng cốt, vừa là phối hợp liên ngành với các nhà Vật lý thiên văn học là Miller và Nichol - hai trong số các nhà Vật lý thiên văn học được nhắc đến ở trên. Ban đầu, họ xem xét các vấn đề về phát hiện tín hiệu trong vùng ảnh điểm lớn. Sau đó, phối hợp lại với nhau, các nhà Vật lý nhận ra rằng cách tiếp cận này có thể áp dụng vào các dấu hiệu dao động âm thanh đặc trưng. Đó là lí do mà bài viết ra đời trên Tạp chí Science.

Các nhà Vật lý thiên văn học Miller và Nichol cho biết: khi thảo luận với cộng đồng Vật lý học về công trình nghiên cứu này, họ đã nhận được rất nhiều quan tâm về phương pháp tiếp cận theo Tỷ lệ phát hiện sai. Giáo sư vật lý Bob Nichol thuộc trường Đại học Carnegie Mellon có viết: "Cá nhân tôi muốn nhấn mạnh về mối quan hệ cộng sinh không ngừng phát triển giữa các nhà Thống kê và các nhà Vật lý thiên văn tại trường Đại học Carnegie Mellon. Rõ ràng là có những vấn đề cốt lõi thường thấy mà các nhà nghiên cứu trong cả hai lĩnh vực đều quan tâm, ví dụ như Ứng dụng của Tỷ lệ phát hiện sai vào các vấn đề thuộc Vật lý thiên văn".

Trên thực tế, các nhà Vật lý thiên văn học đánh giá cao vẻ đẹp Toán học trong ngành Thống kê, trong khi các nhà Thống kê lại muốn được hỗ trợ cho Vật lý thiên văn tìm hiểu về vũ trụ. Bên cạnh các dự án liên ngành như vậy, sự hợp tác giữa Thống kê học và Vật lý thiên văn học cũng đẩy mạnh các nghiên cứu mới riêng biệt trong mỗi ngành. Tóm lại, sự hợp tác đa phương này đã thúc đẩy mạnh mẽ cả các nghiên cứu chung và các nghiên cứu riêng mới mẻ trong các lĩnh vực khoa học. Có thể nói, đây chính là một sự kết hợp hoàn hảo!

(Còn tiếp)

Quỳnh Trang (dịch) - Đoàn Dũng (hiệu đính)

Nguồn: A Report on the Future of Statistics http://ww.biostat.jhsph.edu/...