SO SÁNH SAS, SPSS VÀ STATA

Lê Đỗ Mạch
Viện khoa học thống kê

Hiện nay có ba bộ chương trình chuyên dụng phục vụ cho xử lý và phân tích số liệu thống kê rất thông dụng trên thế giới, đó là SAS, SPSS và STATA. Các chương trình này không những được giảng dạy trong các trường đại học mà còn là những công cụ không thể thiếu được đối với các nhà thống kê và các nghiên cứu quan sát thống kê ở nhiều lĩnh vực khác nhau. Trong số ba bộ chương trình thì SAS là chương trình lớn nhất và mạnh nhất nhưng lại đắt nhất, nên trong giai đoạn hiện nay ít được phổ biến ở nước ta; còn hai bộ chương trình SPSS và STATA nhiều người biết và đang sử dụng trong nghiên cứu thống kê từ đầu những năm 1990.

Vậy, sự khác nhau của STATA với SAS và SPSS là như thế nào? và bộ chương trình nào là tốt nhất. Mỗi bộ chương trình đều có đặc trưng riêng của nó, những điểm mạnh và yếu của nó. Bài viết này sẽ

tóm tắt đặc trưng, điểm mạnh và điểm yếu riêng của từng bộ chương trình trên cả bốn phương diện:

1. Về sử dụng

SAS là bộ chương trình mà nhiều người sử dụng có trình độ cao ưa thích bởi sức mạnh và khả năng lập trình của nó. Do SAS là một bộ chương trình mạnh như vậy nên khó học nhất. Để sử dụng SAS, ta phải viết chương trình để thao tác dữ liệu và thực hiện các phân tích dữ liệu của mình. Nếu chương trình mắc lỗi, cái khó là phải biết tìm lỗi ở đâu và cách sửa thế nào.

SPSS là một bộ chương trình mà nhiều người sử dụng ưa thích do nó rất dễ sử dụng. SPSS có một giao diện giữa người và máy cho phép sử dụng các menu thả xuống để chọn các lệnh thực hiện. Khi thực hiện một phân tích chỉ đơn giản chọn thủ tục cần thiết và chọn các biến phân tích và bấm OK là có kết quả ngay trên màn hình để xem xét. SPSS cũng có một ngôn ngữ cú pháp có thể học bằng cách dán cú pháp lệnh vào cửa sổ cú pháp từ một lệnh vừa chọn và thực hiện, nhưng nói chung khá phức tạp và không trực giác.

STATA là một bộ chương trình mà nhiều người mới bắt đầu và sử dụng mạnh đều ưa thích vì nó vừa dễ học có nhiều khả năng. STATA sử dụng các lệnh trực tiếp, có thể vào mỗi lệnh ở một thời điểm để thực hiện (chế độ này được người mới bắt đầu ưa thích) hoặc có thể soạn thảo thành một chương trình bao gồm nhiều lệnh cho một nhiệm vụ và thực hiện cùng một lúc. Thậm chí nếu mắc lỗi trong chương trình thì có thể nhận biết và sửa chữa dễ dàng.

2. Về quản lý dữ liệu

SAS rất mạnh trong lĩnh vực quản lý dữ liệu, cho phép người sử dụng thao tác dữ liệu hầu như với mọi cách có thể. SAS cũng đưa vào thủ tục Proc sql cho phép thực hiện mọi câu hỏi Sql (Structured query language) trên file dữ liệu. Tuy nhiên phải mất nhiều thời gian để học và hiểu được cách quản lý dữ liệu của SAS và nhiều nhiệm vụ quản lý phức tạp của nó lại được thực hiện bằng những lệnh đơn giản trong SPSS và STATA. Thay vào đó, SAS có thể làm việc với nhiều file dữ liệu cùng một lúc; điều này giảm đi tính phức tạp trong chuẩn bị dữ liệu đối với những nhiệm vụ phân tích đòi hỏi phải làm việc với nhiều file dữ liệu cùng một lúc. Trong khi đó mỗi thời điểm STATA hoặc SPSS chỉ làm việc được với một file dữ liệu. SAS có thể quản lý những file dữ liệu khổng lồ lên đến 32.768 biến và số lượng bản ghi là rất lớn chỉ phụ thuộc vào kích cỡ của đĩa cứng. Ưu điểm này có thể làm đơn giản hoá trong khi tổ chức xử lý và phân tích trên một khối lượng rất lớn dữ liệu vì dữ liệu chỉ chứa trong một file.

SPSS có một bộ soạn thảo dữ liệu tương tự như excel, bộ soạn thảo cho phép vào các dữ liệu và mô tả các thuộc tính của chúng, tuy nhiên SPSS không có những công cụ quản lý dữ liệu thật mạnh (mặc dù SPSS phiên bản 11 có thêm các lệnh chuyển cấu trúc dữ liệu theo chiều ngang thành cấu trúc dữ liệu theo chiều dọc và ngược lại....). SPSS xử lý mỗi file dữ liệu ở một thời điểm và không phải là rất mạnh khi thực hiện các nhiệm vụ phân tích cần làm việc với nhiều file dữ liệu cùng một lúc. Các file dữ liệu có thể có đến 4096 biến và số lượng bản ghi chỉ bị giới hạn trong dung lượng của đĩa cứng.

STATA hoàn toàn không có khả năng quản lý dữ liệu mạnh như SAS, nhưng các lệnh quản lý dữ liệu của nó vẫn có nhiều sức mạnh, lại rất đơn giản. Chúng cho phép thực hiện các thao tác phức tạp về dữ liệu một cách dễ dàng. Tuy nhiên, mỗi thời điểm STATA chỉ làm việc được với một file dữ liệu, vì vậy những nhiệm vụ xử lý cần nhiều file dữ liệu cùng một lúc đối với STATA là phức tạp hơn. Với việc đưa vào bộ giải phóng STATA/Se, số lượng biến có thể có đến 32.768 biến trong một file dữ liệu STATA, và kích cỡ của file cũng chỉ phụ thuộc vào dung lượng đĩa cứng.

3. Về phân tích thống kê

SAS, SPSS, STATA cùng tính toán những thống kê mô tả và thực hiện những phân tích thống kê chung nhất như hồi qui, hồi qui logistic, phân tích tồn tại, phân tích phương sai, phân tích nhân tố, và phân tích nhiều chiều.

Trước hết xin bàn về khả năng tổng hợp số liệu (tính toán các thống kê mô tả). Một trong những công việc thường xuyên phải làm đối với cán bộ nghiệp vụ thống kê là tổng hợp số liệu theo các biểu bảng đã thiết kế trước đối với số liệu thu được. Nếu ai đã từng sử dụng SPSS và STATA, đều thấy rằng khả năng lập các biểu bảng số liệu tổng hợp, các báo cáo thống kê trên tập số liệu cơ sở trong SPSS là hết sức đa dạng và linh hoạt với nhiều chiều phân tổ khác nhau và dễ dàng thực hiện không phải lập trình. Các bảng biểu, các báo cáo được trình bày đẹp, chất lượng cao được hiện trên cửa sổ, có thể tiếp tục hiệu chỉnh, in ra hoặc chuyển sang các tài liệu khác. Đây là một ưu điểm nổi bật của SPSS, vì để lập trình tạo ra một biểu bảng như ý là một công việc hết sức tỉ mẩn và nặng nhọc.

Đối với phân tích thống kê, sức mạnh lớn nhất của SAS có thể tìm thấy trong phân tích ANOVA, phân tích mô hình hỗn hợp và phân tích nhiều chiều, trong khi nó lại tỏ ra yếu với hồi qui logistic kiểu thứ tự và kiểu phạm trù (vì các lệnh này là đặc biệt khó) và các phương pháp ước lượng mạnh. Nó cũng có hỗ trợ một ít cho phân tích dữ liệu theo lược đồ mẫu, nhưng lại hạn chế hơn so với STATA.

Sức mạnh lớn nhất của SPSS là lĩnh vực phân tích phương sai (SPSS cho phép thực hiện nhiều loại kiểm định tác động riêng biệt) và phân tích nhiều chiều (thí dụ phân tích phương sai nhiều chiều, phân tích nhân tố, phân tích nhóm tổ). SPSS phiên bản 11 còn bổ sung thêm một số khả năng phân tích các mô hình hỗn hợp. Cái yếu nhất của SPSS là khả năng xử lý đối với những vấn đề ước lượng phức tạp và do đó khó đưa ra được các ước lượng sai số đối với các ước lượng này. SPSS cũng không hỗ trợ các công cụ phân tích dữ liệu theo lược đồ mẫu.

Sức mạnh lớn nhất của STATA là hồi qui (rất dễ sử dụng các công cụ đoán nhận hồi qui), hồi qui logistic (những bổ sung mới làm đơn giản hoá việc giải thích kết quả hồi qui logistic, còn hồi qui logistic thứ tự và hồi qui logistic phạm trù là rất dễ thực hiện). STATA cũng có nhiều phương pháp ước lượng mạnh rất dễ sử dụng, bao gồm cả hồi qui mạnh và hồi qui với sai số chuẩn mạnh, và nhiều lệnh ước lượng khác kèm theo sai số chuẩn mạnh. STATA cũng trội hơn về lĩnh vực phân tích dữ liệu theo lược đồ mẫu, cho khả năng áp dụng chúng trong phân tích số liệu điều tra bởi các công cụ hồi qui, hồi qui logistic, hồi qui poisson, hồi qui probit,... Điểm yếu nhất là khả năng phân tích phương sai và phân tích nhiều chiều truyền thống

như phân tích phương sai nhiều chiều, phân tích nhóm tổ.

4. Về vẽ đồ thị

SAS có các công cụ vẽ đồ thị mạnh nhất (SAS/Graph) so với hai bộ chương trình còn lại. Để sử dụng SAS/Graph phải yêu cầu có chuyên môn và không đơn giản. Có thể tạo ra các đồ thị đa dạng bằng cú pháp, tuy nhiên SAS 8 có giao diện giữa người và máy để tạo ra các đồ thị, nhưng không dễ sử dụng như SPSS.

SPSS có một giao diện giữa người và máy rất đơn giản để tạo ra các đồ thị và khi đã tạo được một đồ thị, nhờ giao diện này mà người sử dụng có thể tuỳ ý hiệu chỉnh đồ thị cũng như hoàn thiện chúng. Các đồ thị có chất lượng rất cao và có thể dán vào các tài liệu khác, thí dụ như Word hoặc Powerpoint. SPSS có ngôn ngữ cú pháp để tạo ra các đồ thị, nhưng nhiều điểm trong giao diện tạo đồ thị lại không sẵn sàng trong ngôn ngữ cú pháp. Ngôn ngữ cú pháp của SPSS phức tạp hơn so với STATA, nhưng lại có phần đơn giản hơn, ít mạnh hơn SAS.

Giống như SPSS, các đồ thị STATA có thể tạo ra bằng sử dụng lệnh hoặc giao diện giữa người sử dụng và máy (STATA 8), nhưng khác hơn SPSS ở chỗ các đồ thị của STATA không thể hiệu đính bằng bộ hiệu đính đồ thị. Cú pháp của các lệnh đồ thị là dễ sử dụng nhất trong số ba bộ chương trình và cũng là mạnh nhất. Các đồ thị STATA có chất lượng cao và chất lượng xuất bản cũng cao. Thêm vào đó các đồ thị STATA còn có chức năng bổ sung cho phân tích thống kê, thí dụ như có nhiều lệnh làm đơn giản hoá việc tạo ra các đồ thị chẩn đoán hồi qui.

Tóm lại, SAS là một bộ chương trình hướng tới những người sử dụng có trình độ cao, khó học và nhất là lúc ban đầu. Tuy nhiên những người sử dụng mạnh thích sức mạnh quản lý dữ liệu và khả năng làm việc cùng một lúc với nhiều file dữ liệu lớn của SAS.

SPSS nhắm vào mục tiêu dễ sử dụng, khẩu hiệu của họ là thực sự làm, thực sự dễ, và mục tiêu này đã thành công. Nhưng nếu ta dự định sử dụng SPSS như một người sử dụng mạnh, thì nó có thể không đáp ứng được yêu cầu. SPSS mạnh về lĩnh vực đồ thị và lập biểu bảng, báo cáo tổng hợp số liệu, nhưng lại yếu hơn về một số thủ tục thống kê như phương pháp ước lượng mạnh và thiếu vắng phương pháp phân tích dữ liệu theo lược đồ mẫu.

STATA cho một kết hợp tốt giữa dễ sử dụng và sức mạnh. Trong khi STATA dễ học và cũng có những công cụ mạnh về quản lý dữ liệu, nhưng cũng như đã nêu trong phần phân tích có một số thủ tục thống kê cũng bị cắt bỏ. Trong STATA khả năng tải các chương trình phát triển bởi những người sử dụng khác về là dễ dàng và đồng thời có khả năng tạo ra các chương trình riêng của người sử dụng, để chúng trở thành một bộ phận của STATA.

Như một tổng thể, các chương trình SAS, SPSS, STATA hình thành một tập hợp các công cụ đa dạng trên một phạm vi rộng dùng trong phân tích thống kê. Với chương trình Stat/Transfer, rất dễ dàng chuyển các file dữ liệu từ bộ chương trình này sang bộ chương trình khác một cách nhanh chóng. Do đó, để tận dụng được thế mạnh của từng bộ chương trình khi phân tích số liệu, chúng ta có thể chuyển từ bộ chương trình này sang bộ chương trình kia để phân tích, điều đó phụ thuộc vào bản chất của vấn đề đang nghiên cứu. Thí dụ nếu trong khi đang thực

hiện phân tích cần sử dụng mô hình hỗn hợp thì có thể chọn SAS, nhưng nếu làm hồi qui logistic có thể chọn STATA, còn nếu phân tích phương sai hoặc nhóm tổ có thể chọn SPSS và để tổng hợp số liệu theo biểu bảng thì dùng SPSS■