1
Lời nói đầu

Trái với quan điểm của nhiều người, thống kê là một bộ môn khoa học: *Khoa học thống kê* (Statistical Science). Các phương pháp phân tích dù dựa vào nền tảng của toán học và xác suất, nhưng đó chỉ là phần "kĩ thuật", phần quan trọng hơn là thiết kế nghiên cứu và diễn dịch ý nghĩa dữ liệu. Người làm thống kê, do đó, không chỉ là người đơn thuần làm phân tích dữ liệu, mà phải là một nhà khoa học, một nhà suy nghĩ ("thinker") về nghiên cứu khoa học. Chính vì thế, mà khoa học thống kê đóng một vai trò cực kì quan trọng, một vai trò không thể thiếu được trong các công trình nghiên cứu khoa học, nhất là khoa học thực nghiệm. Có thể nói rằng ngày nay, nếu không có thống kê thì các thử nghiệm gen với triệu triệu số liệu chỉ là những con số vô hồn, vô nghĩa.

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 có ý nghĩa khoa học gì cả. Chính vì thế thế mà 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 tạp san y học có uy tín càng cao yêu cầu về phân tích thống kê càng nặng. Xin nhắc lại để nhấn mạnh: không có phần phân tích thống kê, bài báo không có ý nghĩa khoa học.

Một trong những phát triển quan trọng nhất trong khoa học thống kê là ứng dụng máy tính cho phân tích và tính toán thống kê. Có thể nói không ngoa rằng không có máy tính, khoa học thống kê vẫn chỉ là một khoa học buồn tẻ khô khan, với những công thức rắc rối mà thiếu tính ứng dụng vào thực tế. Máy tính đã giúp khoa học thống kê làm một cuộc cách mạng lớn nhất trong lịch sử của bộ môn: đó là đưa khoa học thống kê vào thực tế, giải quyết các vấn đề gai góc nhất và góp phần làm phát triển khoa học thực nghiệm.

Người viết còn nhớ hơn 20 năm về trước khi còn là một sinh viên theo học chương trình thạc sĩ thống kê ở Úc, một vị giáo sư khả kính kể một câu chuyện về nhà thống kê danh tiếng người Mĩ, Fred Mosteller, nhận được một hợp đồng nghiên cứu từ Bộ Quốc phòng Mĩ để cải tiến độ chính xác của vũ khí Mĩ vào thời Thế chiến thứ II, mà trong đó ông phải giải một bài toán thống kê gồm khoảng 30 thông số. Ông phải mướn 20 sinh viên sau đại học làm việc này: 10 sinh viên chỉ việc suốt ngày tính toán bằng tay; còn 10 sinh viên khác kiểm tra lại tính toán của 10 sinh viên kia. Công việc kéo dài gần một tháng trời. Ngày nay, với một máy tính cá nhân (personal computer) khiêm tốn, phân tích thống kê đó có thể giải trong vòng trên dưới 1 giây.

Nhưng nếu máy tính mà không có phần mềm thì máy tính cũng chỉ là một đống sắt hay silicon "vô hồn" và vô dụng. Một phần mềm đã, đang và sẽ làm cách mạng thống kê là R. Phần mềm này được một số nhà nghiên cứu thống kê và khoa học trên thế giới phát triển và hoàn thiện trong khoảng 10 năm qua để sử dụng cho việc học tập, giảng dạy và nghiên cứu. Cuốn sách này sẽ giới thiệu bạn đọc cách sử dụng R cho phân tích thống kê và đồ thị.

Tại sao R? Trước đây, các phần mềm dùng cho phân tích thống kê đã được phát triển và khá thông dụng. Những phần mềm nổi tiếng từ thời "xa xưa" như MINITAB, BMD-P đến những phần mềm tương đối mới như STATISTICA, SPSS, SAS, STAT, v.v… thường rất đắt tiền (giá cho một đại học có khi lên đến hàng trăm ngàn đô-la hàng năm), một cá nhân hay thậm chí cho một đại học không khả năng mua. Nhưng R đã thay đổi tình trạng này, vì R hoàn toàn miễn phí. Trái với cảm nhận thông thường, miễn phí không có nghĩa là chất lượng kém. Thật vậy, chẳng những hoàn toàn miễn phí, R còn có khả năng làm tất cả (xin nói lại: tất cả), thậm chí còn hơn cả, những phân tích mà các phần mềm thương mại làm. R có thể tải xuống máy tính cá nhân của bất cứ cá nhân nào, bất cứ lúc nào, và bất cứ ở đâu trên thế giới. Chỉ vài phút cài đặt là R có thể đưa vào sử dụng. Chính vì thế mà đại đa số các đại học Tây phương và thế giới càng ngày càng chuyển sang sử dụng R cho học tập, nghiên cứu và giảng dạy. Trong xu hướng đó, cuốn sách này có một mục tiêu khiêm tốn là giới thiệu đến bạn đọc trong nước để kịp thời cập nhật hóa những phát triển về tính toán và phân tích thống kê trên thế giới.

Cuốn sách này được soạn chủ yếu cho sinh viên đại học và các nhà nghiên cứu khoa học, những người cần một phần mềm để học thống kê, để phân tích số liệu, hay vẽ đồ thị từ số liệu khoa học. Cuốn sách này không phải là sách giáo khoa về lí thuyết thống kê, hay nhằm chỉ bạn đọc cách làm phân tích thống kê, nhưng sẽ giúp bạn đọc làm phân tích thống kê hữu hiệu hơn và hào hứng hơn. Mục đích chính của tôi là cung cấp cho bạn đọc những kiến thức cơ bản về thống kê, và cách ứng dụng R cho giải quyết vấn đề, và qua đó làm nền tảng để bạn đọc tìm hiểu hay phát triển thêm R.

Tôi cho rằng, cũng như bất cứ ngành nghề nào, cách học phân tích thống kê hay nhất là tự mình làm phân tích. Vì thế, sách này được viết với rất nhiều ví dụ và dữ liệu thực. Bạn đọc có thể vừa đọc sách, vừa làm theo những chỉ dẫn trong sách (bằng cách gõ các lệnh vào máy tính) và sẽ thấy hào hứng hơn. Nếu bạn đọc đã có sẵn một dữ liệu nghiên cứu của chính mình thì việc học tập sẽ hữu hiệu hơn bằng cách ứng dụng ngay những phép tính trong sách. Đối với sinh viên, nếu chưa có số liệu sẵn, các bạn có thể dùng các phương pháp mô phỏng (simulation) để hiểu thống kê hơn.

Khoa học thống kê ở nước ta tương đối còn mới, cho nên một số thuật ngữ chưa được diễn dịch một cách thống nhất và hoàn chỉnh. Vì thế, bạn đọc sẽ thấy đây đó trong sách một vài thuật ngữ "lạ", và trong trường hợp này, tôi cố gắng kèm theo thuật ngữ gốc

tiếng Anh để bạn đọc tham khảo. Ngoài ra, trong phần cuối của sách, tôi có liệt kê các thuật ngữ Anh – Việt đã được đề cập đến trong sách.

Tất cả các dữ liệu sử dụng trong sách này đều có thể tải từ internet xuống máy tính cá nhân, hay có thể truy nhập trực tiếp qua trang web: http://www.ykhoa.net/R.

Tôi hi vọng bạn đọc sẽ tìm thấy trong sách một vài thông tin bổ ích, một vài kĩ thuật hay phép tính có ích cho việc học tập, giảng dạy và nghiên cứu của mình. Nhưng có lẽ chẳng có cuốn sách nào hoàn thiện hay không có thiếu sót; thành ra, nếu bạn đọc phát hiện một sai sót trong sách, xin báo cho tôi biết qua điện thư t.nguyen@garvan.org.au hay rknguyen@gmail.com. Thành thật cám ơn các bạn đọc trước.

Tôi muốn nhân dịp này cám ơn Tiến sĩ Nguyễn Hoàng Dzũng thuộc khoa Hóa, Đại học Bách khoa Thành phố Hồ Chí Minh, người đã gợi ý và giúp đỡ tôi in cuốn sách này ở trong nước. Tôi cám ơn Bác sĩ Nguyễn Đình Nguyên, người đã đọc một phần lớn bản thảo của cuốn sách, góp nhiều ý kiến thiết thực, và đã thiết kế bìa sách. Tôi cũng cám ơn Nhà xuất bản Đại học Bách khoa Thành phố Hồ Chí Minh đã giúp tôi in cuốn sách này.

Bây giờ, tôi mời bạn đọc cùng đi với tôi một "hành trình thống kê" ngắn bằng R.

Sydney, 31 Tháng Ba Năm 2006
Nguyễn Văn Tuấn