# TƯƠNG LAI CỦA THỐNG KÊ HỌC
## (Tiếp theo)

### 4. ỨNG DỤNG THỐNG KÊ TRONG CÁC NGÀNH KHOA HỌC VÀ CÔNG NGHIỆP

Một đặc điểm riêng biệt của Thống kê học chính là có sự tương tác với toàn bộ các ngành khoa học tự nhiên và xã hội cũng như các ngành công nghệ. Trong phần này, chúng ta sẽ lý giải về vai trò của Thống kê học trong việc thu thập thông tin. Thay vì tiến hành một cuộc khảo sát về các lĩnh vực trên diện rộng và chưa đầy đủ về những tác động của ngành Thống kê trong quá khứ, hiện tại và tương lai, nội dung này sẽ tập trung vào khía cạnh quan trọng về mối quan hệ tương tác giữa ngành Thống kê và các ngành khoa học khác.

#### 4.1. Sinh học

Hình thành trên cơ sở các *số liệu thống kê* ngành *nông nghiệp* và di truyền được phát triển trong nửa đầu thế kỷ XX, Thống kê sinh học, Thống kê dịch tễ học và các thử nghiệm lâm sàng ngẫu nhiên là nền tảng cho việc điều trị các căn bệnh của con người một cách có hệ thống, góp phần gia tăng đáng kể tuổi thọ trung bình ở các nước phát triển trong nửa thế kỷ qua.

Những tiến bộ gần đây trong Sinh học phân tử và Di truyền học đã mở ra các lĩnh vực điều tra Thống kê hoàn toàn mới, từ đó sẽ đạt được những tiến bộ nhanh chóng trong việc tìm hiểu quá trình sống cơ bản ở cấp độ phân tử trong tương lai gần. Mục tiêu dài hạn của nghiên cứu này là áp dụng kiến thức về các quá trình phân tử vào các sinh vật và quần thể hoàn chỉnh. Bên cạnh đó, nghiên cứu còn nhằm cải thiện các biện pháp điều trị y tế cho từng cá thể (chẳng hạn như đưa ra các phương pháp điều trị phù hợp với cấu trúc di truyền của từng cá thể), giảm tình trạng suy dinh dưỡng và chết đói bằng cách cải thiện các loài cây trồng nông nghiệp và vật nuôi chủ yếu, cải thiện sức khỏe công cộng và tăng cường quốc phòng chống lại khủng bố sinh học.

Để tóm lược những bước phát triển mới trong nghiên cứu Sinh học, ta có thể xem xét bốn lĩnh vực mà trong đó các phương pháp thống kê và tính toán đã, đang và sẽ tiếp tục được ứng dụng và đóng vai trò quan trọng:

• *Phân tích trình tự phân tử Sinh học và Di truyền học*

Đây là môn khoa học dựa trên phân tích trình tự DNA (khối gen cấu trúc) và các chuỗi Axit Amin (khối Protein cấu trúc), cũng như dựa trên các hồ sơ toàn cầu về RNA và Protein trong các trạng thái tế bào khác nhau, được sử dụng để tìm ra cấu trúc và sự tiến hóa của các Gen và Protein cũng như các chức năng của chúng trong các quá trình bình thường và bất thường. Lấy ví dụ như việc xác định các vùng kiểm soát được mã hóa trong bộ gen điều chỉnh số lượng và điều kiện sản sinh Protein.

• *Dịch tễ học di truyền*

Mục tiêu của Dịch tễ học di truyền là nhằm tìm hiểu về tầm quan trọng tương đối của yếu tố môi trường và di truyền đối với bệnh tật của con người. Ví dụ, lập bản đồ gen liên quan đến việc sử dụng các

bản đồ dấu hiệu phân tử trong bộ gen của một loài thực vật hoặc động vật cụ thể để xác định vị trí các gen góp phần vào kiểu hình mà ta quan tâm. Đây thường là bước khởi đầu để hiểu rõ hơn và điều trị các bệnh ở thực vật và động vật - vấn đề mà yếu tố di truyền đóng một vai trò hết sức quan trọng.

• *Tiến hóa, Di truyền học quần thể và Sinh thái học*

Đây là các mảng nghiên cứu những thay đổi xảy ra ở cấp độ quần thể ở thực vật và động vật phản ứng với những thay đổi đột biến ngẫu nhiên trong gen của quần thể và những thay đổi trong môi trường của chúng. Mặc dù ban đầu hướng tới việc nghiên cứu mối quan hệ tiến hóa (ví dụ như, bằng chứng ủng hộ giả thuyết *con người hiện đại* có nguồn gốc tiến hóa từ *Phi châu*), Di truyền học quần thể ngày càng được áp dụng để nghiên cứu về sự tiến hóa của vi khuẩn và virus (nhằm tìm ra các loại vắc xin và thuốc thích hợp) và sự tiến hóa của Protein trong các loài thực vật và động vật khác nhau (để hiểu cấu trúc và chức năng của Protein, cần xác định các bộ phận của các Protein liên quan từ nhiều loài khác nhau còn tồn tại thông qua tiến hóa).

• *Khoa học thần kinh tính toán*

Là mảng sử dụng các *phương pháp hình ảnh* chức năng *thần kinh* hiện đại - Neuroimaging (PET, fMRI) để nghiên cứu các chức năng của hệ thần kinh. Điều này đặt ra những câu hỏi ở cả cấp độ số lượng nhỏ các tế bào thần kinh tương tác và cấp độ toàn bộ bộ não:

Các bộ phận nào của não bộ được hoạt hóa theo điều kiện nào? Cấu trúc và/hoặc chức năng bộ não của những người bình thường và bệnh nhân tâm thần khác nhau như thế nào? Làm thế nào chúng ta có thể áp dụng kiến thức này trong chẩn đoán và điều trị?

***Các phương pháp thống kê và tính toán***

Như một hệ quả của nhiều vấn đề khoa học đa dạng, một tập hợp mở rộng các phương pháp thống kê, xác suất và tính toán đã được chứng minh là rất hữu ích. Một số phương pháp đã tự chứng minh trong một số lĩnh vực, trong khi những phương pháp khác lại có các ứng dụng chuyên biệt hơn.

Các quy trình ngẫu nhiên, từ chuỗi Markov hữu hạn đến các quá trình điểm và các trường ngẫu nhiên Gauss rất hữu ích trên toàn bộ phạm vi các vấn đề. Do rất nhiều các dữ liệu được tạo ra [ví dụ như, thiết bị microarray cho phép biểu hiện hàng chục ngàn gen trong một mẫu cá thể, hoặc dữ liệu từ khoảng 1.000 dấu hiệu (trong tương lai có thể đạt mức 100.000) phân bố trên toàn bộ gen của hàng ngàn cá thể], những thách thức trong so sánh đa giả thuyết thường xuyên phát sinh trong các lĩnh vực này.

Mô hình Markov ẩn và chuỗi Markov phương pháp Monte Carlo cung cấp các thuật toán tính toán quan trọng để tính toán và tối đa hóa các chức năng có thể. Một số phương pháp thống kê là các phương pháp cổ điển (ví dụ như, các thành phần chủ yếu, phân tích khả năng), nhưng các phương pháp này thậm chí có thể yêu cầu sự thích ứng (các đường chính, phân tích khả năng các quá trình ngẫu nhiên) để đối phó với số lượng dữ liệu lớn từ các thí nghiệm Sinh học hiện đại. Gần đây, các phương pháp khác (mô hình Markov ẩn và chuỗi Markov phương pháp Monte Carlo) đã tương đối phát triển cùng với công nghệ máy tính hiện đại cần thiết để thực hiện các phương pháp này.

Một tính năng phổ biến của tất cả các nỗ lực được mô tả ở trên là số lượng, tính phức tạp và tính biến thiên của dữ liệu, với kết quả tính toán (thường bao gồm cả đồ họa) là một khía cạnh quan trọng

trong việc thực hiện ý tưởng. Trên quan điểm các cơ sở Toán học và Điện toán đa dạng của các nhà khoa học tham gia vào nghiên cứu Sinh học, điều quan trọng là thuật toán tính toán được tạo ra nhằm dễ sử dụng nhất có thể. Điều này có thể cần đến sự hỗ trợ cho các chuyên gia để cung cấp "giao diện phía trước" và tài liệu cần thiết để các nhà khoa học trong phòng thí nghiệm có thể sử dụng dễ dàng và chính xác các công cụ được các nhà Thống kê phát triển.

Tóm lại, nhiều dữ liệu được tạo ra từ các thí nghiệm Sinh học hiện đại. Đồng thời, tính biến đổi trong con người nhằm phản ứng lại các can thiệp y tế tạo ra nhu cầu ngày càng tăng đối với các nhà Thống kê- những người có thể kết hợp với các nhà Sinh học và đưa ra những phương pháp mới để hướng dẫn thiết kế thí nghiệm và phân tích dữ liệu Sinh học.

**4.2. Kỹ thuật và Công nghiệp**

Các khái niệm và các phương pháp thống kê đã đóng một vai trò hết sức quan trọng trong sự phát triển của ngành công nghiệp trong thế kỷ qua. Lần lượt các ứng dụng trong kỹ thuật và công nghiệp đã trở thành chất xúc tác chính trong việc nghiên cứu lý thuyết và phương pháp luận thống kê. Sự phong phú và đa dạng của những vấn đề này có ảnh hưởng lớn đến sự phát triển của ngành Thống kê.

Phần lớn các hoạt động trước đây của ngành Thống kê bị chi phối bởi nhu cầu của các ngành nông nghiệp, công nghiệp sản xuất và công nghiệp quốc phòng. Tuy nhiên, trong những năm gần đây, phạm vi đã mở rộng đáng kể sang lĩnh vực kinh doanh và tài chính, công nghệ phần mềm cũng như các ngành công nghiệp dịch vụ và y tế. Những ứng dụng trong các lĩnh vực này bao gồm tính điểm tín dụng, hồ sơ khách hàng, thiết kế đường cao tốc và các loại xe thông minh, thương mại điện tử, phát hiện gian lận, giám sát mạng và cuối cùng là chất lượng và độ tin cậy của phần mềm.

Cạnh tranh toàn cầu và những mong đợi của khách hàng ngày càng tăng đang làm biến đổi môi trường hoạt động của các công ty. Những thay đổi này có ý nghĩa quan trọng đối với các hướng nghiên cứu trong Thống kê. Sau đây là mô tả ngắn gọn về bốn ví dụ khái quát:

***4.2.1. Các bộ dữ liệu lớn với cấu trúc phức tạp***

Chủ đề này đề cập sơ qua tất cả các lĩnh vực của kinh doanh và công nghiệp (cũng như các khu vực khác được thảo luận trong báo cáo này). Các quá trình kinh doanh và sản xuất đang trở nên ngày càng phức tạp. Do đó, hơn bao giờ hết, các kỹ sư và các nhà quản lý cần có các dữ liệu liên quan để định hướng cho việc ra quyết định.

Đồng thời, những tiến bộ trong các công nghệ cảm biến và thu thập dữ liệu đã tạo điều kiện cho việc thu thập thêm nhiều dữ liệu. Những dữ liệu này thường có cấu trúc phức tạp dưới hình thức chuỗi thời gian, quy trình không gian, văn bản, hình ảnh, kích thước đa chiều với cấu trúc phân cấp... Thu thập, mô hình hóa và phân tích những dữ liệu này là những thách thức nghiên cứu vô cùng khó khăn.

Ví dụ như giám sát, chẩn đoán và cải thiện các quy trình sản xuất tiên tiến đòi hỏi các phương pháp mới trong việc nén dữ liệu và khai thác tính năng, phát triển chẩn đoán thông minh và kiểm soát quá trình thời gian thực. Những vấn đề này cũng liên quan đến các nội dung có cùng bản chất như các xu hướng lựa chọn, tính toán, khả năng mở rộng của các thuật toán và trực quan.

#### *4.2.2. Các mô hình tính toán quy mô lớn*

Các mô hình tính toán và mô phỏng ngày càng được sử dụng thường xuyên hơn trong nhiều lĩnh vực ứng dụng. Trong các ngành công nghiệp sản xuất, các yếu tố thị trường cạnh tranh và áp lực đồng thời để giảm chu kỳ phát triển sản phẩm đã làm giảm quy trình kiểm tra thực tế, đồng thời tăng cường sử dụng thiết kế máy tính hỗ trợ và các phương pháp kỹ thuật. Phân tích phần tử hữu hạn và các kỹ thuật khác cũng được sử dụng rộng rãi trong ngành công nghiệp ô tô để thiết kế và tối ưu hóa sản phẩm.

Các xu hướng tương tự xuất hiện trong các ngành công nghiệp sản xuất bán dẫn, máy bay, ngành quốc phòng và các ngành công nghiệp khác. Các mô hình tính toán đa chiều liên quan đến hàng trăm và thậm chí hàng nghìn các thông số và các biến số thiết kế. Một đánh giá chức năng đơn có thể chỉ mất vài ngày để thực hiện dựa trên các nền tảng máy tính cao cấp.

Thử nghiệm, phân tích, trực quan và xác nhận bằng cách sử dụng các mô hình tính toán quy mô lớn làm phát sinh một loạt các thách thức thống kê. Các thách thức này bao gồm: (a) sự phát triển của thiết kế thí nghiệm xấp xỉ và khám phá các bề mặt phản hồi trong không gian đa chiều; (b) kết hợp ngẫu nhiên và không chắc chắn trong các thông số thiết kế và các đặc tính vật liệu vào mô hình tính toán; (c) mô hình hóa, sàng lọc, dự báo và tối ưu hóa.

#### *4.2.3. Mức độ tin cậy và an toàn*

Thiết kế, phát triển và chế tạo các sản phẩm có độ tin cậy cao, đáp ứng mục tiêu an toàn và mục tiêu môi trường cũng có nghĩa là ngành công nghiệp phải đối mặt với những thách thức mới. Sự tập trung thông thường vào độ tin cậy đã xuất hiện trong việc thu thập và phân tích các số liệu "thời gian đến sai hỏng". Điều này đặt ra những khó khăn trong các ứng dụng có độ tin cậy cao với một vài thất bại và kiểm duyệt ở mức độ cao.

May mắn là các tiến bộ trong công nghệ cảm biến đã tạo ra khả năng thu thập nhiều số liệu hơn trong đo lường mức suy thoái và hiệu suất liên quan cũng như các hệ thống và thành phần. Trong khi những dữ liệu này là một nguồn phong phú cung cấp các thông tin đáng tin cậy thì chỉ có rất ít các mô hình và phương pháp nhằm phân tích dữ liệu suy thoái và nhằm kết hợp các mô hình này với phương pháp physics-of-failure (thất bại vật lý) để ước tính, dự báo và duy trì độ tin cậy hiệu quả. Phân tích suy thoái và dự báo device-level failure (thất bại cấp độ thiết bị) là một phần không thể tách rời của công tác dự báo bảo trì trong các hệ thống đắt tiền và có độ tin cậy cao.

Ngoài ra còn có một lượng lớn dữ liệu field-performance có sẵn từ công tác bảo hành và bảo trì các cơ sở dữ liệu. Khai thác dữ liệu về các tín hiệu, các vấn đề quy trình và sử dụng chúng để cải tiến quy trình phải được tập trung một cách nghiêm túc.

#### *4.2.4. Công nghệ phần mềm*

Đây vẫn còn là một lĩnh vực tương đối mới so với các ngành truyền thống của kỹ thuật. Tầm quan trọng của Công nghệ phần mềm đối với Quốc gia được nhấn mạnh bởi sự phụ thuộc ngày càng tăng của nền kinh tế Mỹ và quốc phòng quốc gia vào các phần mềm chất lượng cao và tối quan trọng (Hội đồng Nghiên cứu Quốc gia, năm 1996).

Thống kê có vai trò đáng kể trong công nghệ phần mềm vì dữ liệu là trung tâm để quản lý quá trình phát triển phần mềm. Và các phương pháp thống kê đã cho thấy giá trị của mình trong việc giải quyết một số khía cạnh. Ví dụ như, các suy xét thống kê là rất cần thiết đối với việc xây dựng và sử dụng các số liệu thống kê cho phần mềm một cách hiệu quả. Bên cạnh đó, các ý tưởng thiết kế thử nghiệm là điểm tựa của công nghệ để giảm số lượng các trường hợp cần thiết

phải kiểm tra phần mềm một cách hiệu quả (nhưng không thấu đáo). Hơn nữa, kiểm soát chất lượng thống kê cung cấp cơ sở để phân tích định lượng các bộ phận khác nhau của quy trình phần mềm và cải thiện quy trình liên tục.

**4.3. Khoa học Địa vật lý và Môi trường**

Thuật ngữ "Khoa học địa vật lý và môi trường" bao gồm nhiều lĩnh vực nghiên cứu cụ thể, đặc biệt là nếu như việc nghiên cứu các hiện tượng và quy trình sinh thái được tính vào Khoa học môi trường. Lĩnh vực hoạt động thống kê rộng lớn này không có một lược sử đơn giản cũng như một mô hình phát triển đơn giản. Thật vậy, lịch sử công tác thống kê trong các ngành khoa học Địa vật lý và Môi trường được đan xen với nhiều lĩnh vực như Nông nghiệp, Sinh học, Kỹ thuật dân dụng, Hóa học khí quyển và Sinh thái học.

Việc thu gom và xử lý một lượng lớn dữ liệu là các tính năng trong rất nhiều các thành phần chính của khoa học Địa vật lý và Môi trường như Khí tượng, Hải dương học, Địa chấn học, phát hiện và ảnh hưởng của biến đổi khí hậu, và sự phân tán các chất ô nhiễm trong bầu khí quyển.

Các nhà Thống kê đã tham gia tích cực vào tất cả các lĩnh vực này. Với các phương pháp thống kê tiên tiến bao gồm: các mô hình dữ liệu bốn chiều (spatiotemporal data) phức tạp và các phương pháp tính toán có liên quan, khả năng tương tác trực tiếp giữa các nhà Thống kê và các nhà khoa học Địa vật lý và Môi trường đã tăng lên rất nhiều. Chúng tôi xin đưa ra một vài ví dụ về các thách thức đang được giải quyết.

***4.3.1. Các mô hình quá trình xác định và các mô hình ngẫu nhiên.***

Việc sử dụng song song các mô hình quá trình xác định và các mô hình thống kê là vô cùng quan trọng. Các mô hình quá trình đã thực hiện một cách đặc trưng những khái niệm khoa học cơ bản ví dụ như cân bằng khối lượng trong các thành phần hóa học, giống như một nền tảng và xây dựng các cấu trúc toán học ngắn gọn, súc tích hơn bằng cách bao quát các phương trình mô tả các tương tác vật lý và hóa học, thường là dưới dạng các bộ phương trình vi phân. Mặt khác, các mô hình thống kê thường dựa trên sự mô tả các mô hình dữ liệu quan sát như là một động lực cơ bản để phát triển mô hình. Dần dần, người ta công nhận rằng sự hiểu biết về các quá trình Địa vật lý và Môi trường có thể được nâng cao bằng cách kết hợp ý tưởng từ hai phương pháp tiếp cận mô hình này.

Một phương pháp đã được sử dụng để kết hợp các mô hình quá trình và mô hình thống kê là sử dụng đầu ra của các mô hình xác định như những thông tin đầu vào cho một công thức ngẫu nhiên. Bản báo cáo đầy đủ của NSF đã đưa ra một minh họa chi tiết cho quá trình này, phát sinh trong phân tích chuỗi thời gian hai biến đại diện cho nhiệt độ trung bình phía Bắc và phía Nam bán cầu.

***4.3.2. Số liệu tương quan và xu hướng môi trường***

Nhiều vấn đề môi trường liên quan đến việc phát hiện và dự báo những thay đổi theo thời gian. Ví dụ, một cơ quan giám sát môi trường như Cơ quan Bảo vệ Môi trường sử dụng các ước tính xu hướng để đánh giá sự thành công của các chương trình kiểm soát ô nhiễm và xác định các khu vực cần kiểm soát nghiêm ngặt hơn. Trong mô hình khí hậu xuất hiện một mối quan tâm lớn, đó là phải xác định xem liệu có một xu hướng tổng thể trong dữ liệu, không chỉ đối với các biến số nghiên cứu rộng rãi như nhiệt độ trung bình toàn cầu, mà còn cho rất nhiều các biến số có kết luận không rõ ràng khác.

Đối với các nhà Thống kê, dự báo các thành phần xu hướng với các lỗi tương quan là một vấn đề có tính

lịch sử lâu dài, và hầu hết công việc này liên quan đến sự tương tác đáng kể giữa các nhà Thống kê và các nhà khoa học Địa vật lý và Môi trường. Ví dụ, Giáo sư Sir Gilbert Walker được các nhà Thống kê biết đến với nhiều đóng góp trong phân tích chuỗi thời gian và đặc biệt là các phương trình Yule-Walker. Đồng thời, ông cũng là một nhà Khí tượng học lỗi lạc, người đã dành nhiều thời gian nghiên cứu El Nino-hiện tượng Dao động phương Nam (Southern Oscillation) và những đóng góp này phần lớn là từ một nghiên cứu mà ra.

Sự hợp tác lâu dài giữa các nhà Thống kê và các nhà Địa vật lý đã dẫn đến một loạt các bài báo về phát hiện sự thay đổi trong tầng bình lưu ozone trong đó một lượng lớn các mô hình với các lỗi tương quan đang được xem xét. Tác giả chính của nghiên cứu này - bao gồm phần lớn các bài báo, là một nhà Thống kê nhưng xuất hiện trên các tạp chí không liên quan đến Thống kê chính thống. Đây chính là một minh hoạ tuyệt vời cho việc các lĩnh vực khoa học khác có thể tiếp cận các số liệu thống kê.

***4.3.3. Mô hình hóa thống kê và khái niệm khoa học***

Thông thường, các thay đổi trong hồ sơ số liệu về môi trường được khái niệm trong khuôn khổ *tín hiệu* thống kê *bị nhiễu*. Thật vậy, đây là trường hợp đối với nhiều mô hình đã được thảo luận ở trên, trong đó đưa ra các hình thức khác nhau đối với các thành phần tín hiệu (hoặc hệ thống) và nhiễu (hoặc lỗi) của các mô hình nhằm đại diện tốt hơn cho các quy trình được nghiên cứu.

Một ví dụ trong sự phát triển của tư duy khoa học trong đó có sự hỗ trợ của số liệu thống kê là phân tích chu kỳ quần thể của linh miêu và thỏ rừng Canada. Một loạt các bài báo liên quan đến vấn đề này đã xuất hiện trong *Kỷ yếu của Viện Hàn lâm Khoa học Quốc gia* (Stenseth và đồng sự, năm 1997, 1998, 2004 mục a, b) và Tạp chí *Science*. Tại đây, sự hợp tác giữa các nhà Thống kê và các nhà Sinh thái học đã củng cố lý thuyết khoa học. Một số khái niệm đã được phát triển thông qua công trình này, bao gồm cả mối quan hệ giữa thứ tự hồi quy của mô hình thống kê, sự phức tạp của các hệ thống thông tin phản hồi giữa các loài (trong ví dụ này là linh miêu và thỏ rừng) và ý tưởng các chu kỳ quần thể có thể biểu hiện đồng bộ không gian.

***(Còn tiếp)***

Quỳnh Trang (dịch) - Đoàn Dũng (hiệu đính)
Nguồn: A Report on the Future of Statistics
http://ww.biostat.jhsph.edu/...

**(tiếp theo trang 17)**

**Tài liệu tham khảo:**

1. Akita, Takahiro, Lukman, Rizal Affandi, và Yukino Yamada. 1999. “Bất bình đẳng trong phân phối chi tiêu hộ gia đình ở In-đô-nê-xi-a: Phân tích theo phương pháp phân tổ chỉ số Theil” *Tạp chí kinh tế phát triển* 37, Số.2: 197-221.

2. Anand, Sudhir. 1983. *Bất bình đẳng và nghèo đói ở Ma-Lai-xi-a: Đo lường và phân tổ số liệu*. New York: Oxford University Press.

3. Bourguignon, Francois. 1979. “Các phương pháp phân tổ bất bình đẳng thu nhập.” Tạp chí *Econometrica* 47, Số. 4: 901-20.

4. Deaton, Angus. 1997. *Phân tích các cuộc điều tra hộ gia đình: Phương pháp phân tích vi mô để phục vụ xây dựng chính sách phát triển*. Baltimore và London: Johns Hopkins University Press.