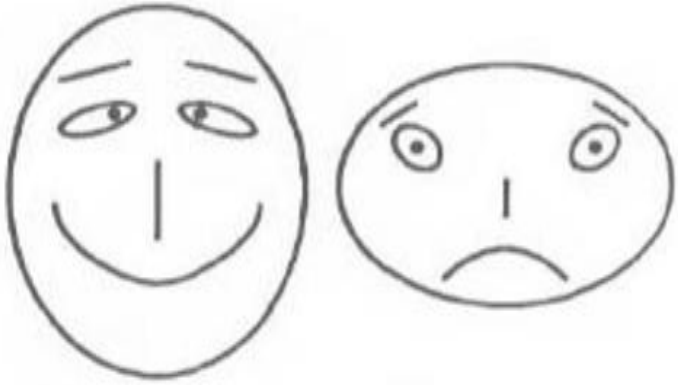
CHƯƠNG 12: KẾT LUẬN

TRÌNH BÀY DỮ LIỆU

Chúng ta đã biết cách trình bày một biến bằng biểu đồ điểm và hai biến bằng việc sử dụng đồ thị phân tán – nhưng làm thế nào để ta có thể cùng lúc biểu thị trên đồ thị mặt phẳng nhiều hơn hai biến? Thực tế có rất nhiều tình huống như vậy, nhưng trong cuốn sách này chúng tôi chỉ xin đề cập tới tới ý tưởng đơn giản của Herman Chernoff, bằng cách sử dụng khuôn mặt, mỗi đặc điểm trên gương mặt được gán cho một biến từ đó ta vẽ được khuôn mặt Chernoff:



x = Độ lệch lông mày

y = Kích thước mắt

z = chiều cao mũi

t = Độ rộng miệng

β = Chiều dài khuôn mặt

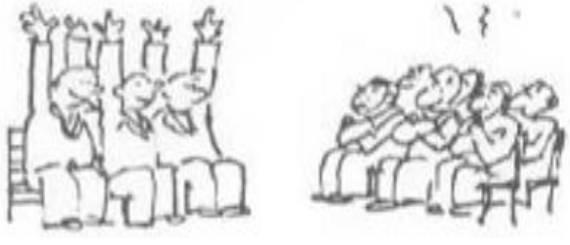
vv...

Phân tích thống kê DỮ LIỆU ĐA BIẾN

Sự phối hợp của các mô hình đa biến giúp phân tích và trình bày dữ liệu theo n chiều, một vài kĩ thuật đa biến:

Phân tích cụm

Phân tích nhằm chia tổng thể thành các nhóm đồng chất. Chẳng hạn, bằng việc phân tích sự bầu cử quốc hội, chúng ta nhận thấy các đại biểu đến từ miền Nam và miền Tây là hai cụm riêng biệt.



Phân tích sự khác biệt

Là một quá trình ngược lại, chẳng hạn Phòng Quản lý đào tạo của một trường đại học đang định tìm kiếm những dữ liệu để có thể đưa ra những dự báo trước liệu một học viên sau khi tốt nghiệp loại Giỏi có thể tiếp tục thành công (đóng góp nhiều cho quỹ cựu sinh viên), hay sẽ là một người không thành công (không phải là công dân tốt trên thế giới và không bao giờ quay trở lại trường).

Phân tích nhân tố

Giúp giải thích dữ liệu đa chiều với một số lượng nhỏ các biến. Một nhà tâm lý học có thể đưa ra bài kiểm tra với 100 câu hỏi, trong khi đó ngầm giả định rằng các câu hỏi chỉ phụ thuộc vào một vài nhân tố: các nhân tố ngoại sinh, nhân tố chủ quan, nhân tố khách quan, v.v... Sau đó, các kết quả kiểm tra sẽ được tổng hợp lại và chỉ sử dụng các điểm số kết hợp của các chiều này.

Các bạn đã biết máy tính có thể giúp ta phân tích và thực hiện các phép tính số học. Nhưng cũng có một số ý tưởng thống kê có sẵn trong máy tính:

Phân tích hình ảnh

Một hình ảnh của máy tính có thể bao gồm hàng nghìn pixel, với mỗi điểm dữ liệu được hiển thị qua một vài pixcel gồm khoảng 16.7 triệu màu. Phân tích hình ảnh thống kê giúp tìm ra ý nghĩa cốt lõi từ "thông tin" giống như bức tranh sau.

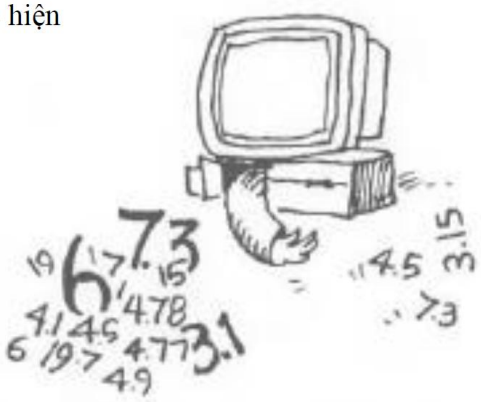
Chọn mẫu lại

Đôi khi không thể tìm được độ lệch chuẩn và giới hạn độ tin cậy. Khi đó, ta tiến hành chọn mẫu lại, một kĩ thuật coi mẫu giống như một tổng thể. Các kĩ thuật này còn có những tên gọi khác như: Randomzation, Jackknife, và Bootstrapping.

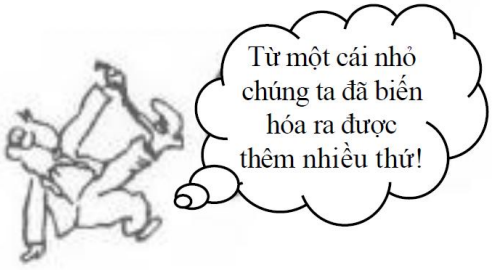
Chọn mẫu lại (tiếp tục)

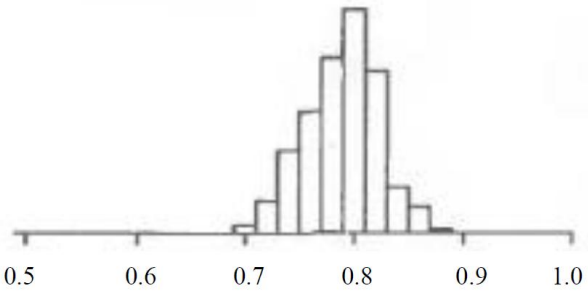
Để tiến hành chọn mẫu lại, chúng ta thực hiện các bước sau:

- Chọn lại mẫu
- Tính các giá trị ước lượng cho mẫu chọn lại
- Lặp lại nhiều lần hai bước trên, tìm khoảng mở rộng của các ước lượng mẫu chọn lại.



Bạn còn nhớ hệ số tương quan r của 92 cặp chiều cao - cân nặng sinh viên trong Chương 11 chứ? Sai số tiêu chuẩn của r bằng bao nhiêu? Máy tính giúp chọn lại 200 mẫu từ 92 điểm dữ liệu, sau đó tính r cho mỗi mẫu mới, và vẽ đồ thị các giá trị r như sau:





Hệ số tương quan của mẫu chọn lại

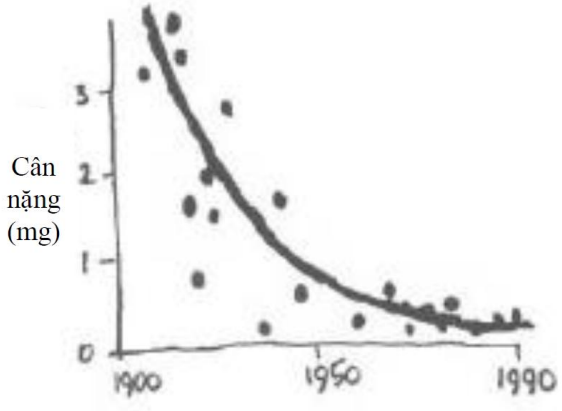
Chú ý khoảng mở rộng của các ước lượng mẫu chọn lại thực tế khá nhỏ.

CHẤT LƯỢNG DỮ LIỆU

Đôi khi một chút lỗi trong việc chọn mẫu, trong phép đo lường, và sai sót trong việc ghi chép dữ liệu có thể làm hỏng quá trình phân tích. R.A Fisher, cha đẻ của nền thống kê hiện đại không chỉ thực hiện công việc thiết kế thí nghiệm và tiến hành phân tích sự sinh sản của động vật mà ông còn trực tiếp chăm sóc và dọn dẹp chuồng nuôi cho chúng, bởi Fisher hiểu rằng chỉ cần mất đi một con vật thì kết quả nghiên cứu cũng sẽ bị ảnh hưởng.

Ngày nay, các nhà thống kê hiện đại, với sự trợ giúp từ các công cụ máy tính, cơ sở dữ liệu, các nguồn trợ cấp từ chính phủ, họ không còn phải trực tiếp tiến hành tất cả các công việc như trước đây.

Nếu bạn biểu diễn khối lượng trung bình mức giảm sinh của chuột dưới tay các nhà thống kê theo thời gian, thì biểu đồ có dạng như sau:



Sự đổi mới

Các phương pháp tốt nhất không phải lúc nào cũng có trong sách. Chẳng hạn như trường hợp một công ty được thuê để ước lượng các thành phần trong rác thải vật liệu, họ gặp phải một vài tình huống rất thú vị mà không hề có trong bất kì cuốn sách nào.

Truyền đạt thông tin

Một phân tích dù có xuất sắc đến mức nào chăng nữa nhưng các kết quả phân tích không thể sát với ngôn ngữ thực tế hay các kết luận chưa đạt được mức độ chắc chắn trong thống kê thì phân tích đó cũng bị coi là thất bại. Đó là lý do tại sao các phương tiện thông tin đại chúng ngày này thường xuyên thông báo về giới hạn sai số trong kết quả bầu cử.

Làm việc nhóm

Trong những mối quan hệ phức tạp của chúng ta, giải pháp của rất nhiều vấn đề đòi hỏi phải có nỗ lực chung của cả nhóm. Các kỹ sư, các nhà thống kê cùng các công nhân của tổ máy phối hợp với nhau để cải thiện chất lượng sản phẩm của họ. Các giáo sư thống kê cùng các nhân viên trợ giúp đang làm việc cùng nhau thiết kế những thử nghiệm ban đầu để tạo ra một phương pháp hiệu quả nhanh chóng.

Biên dịch: Minh Ánh và các nghiên cứu viên, Viện Khoa học Thống kê