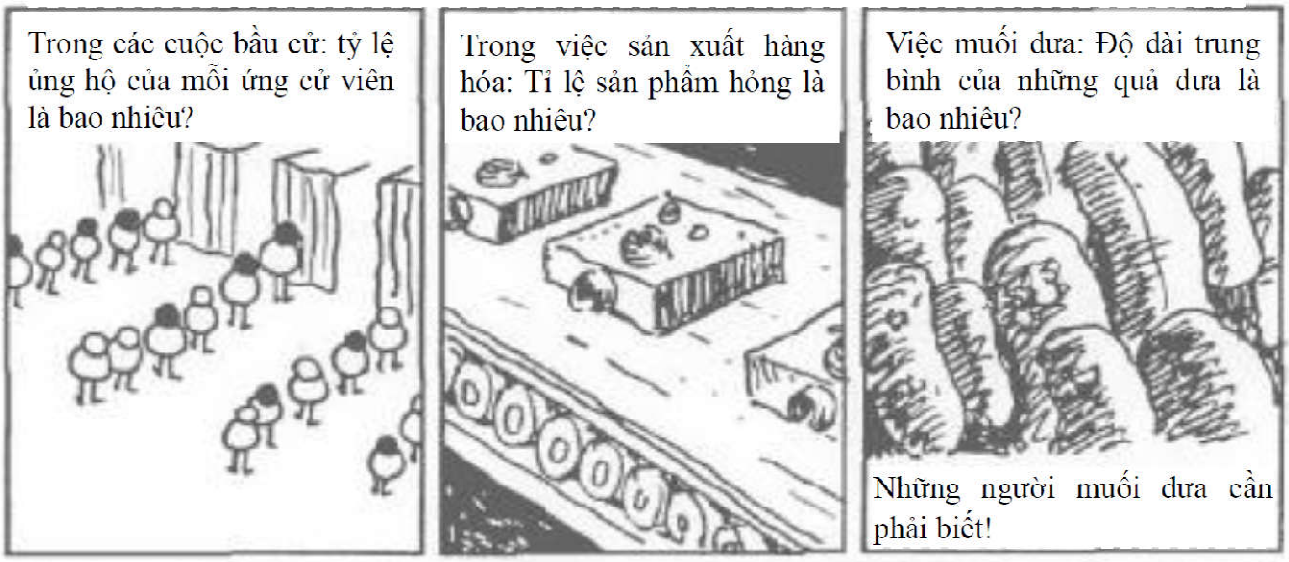
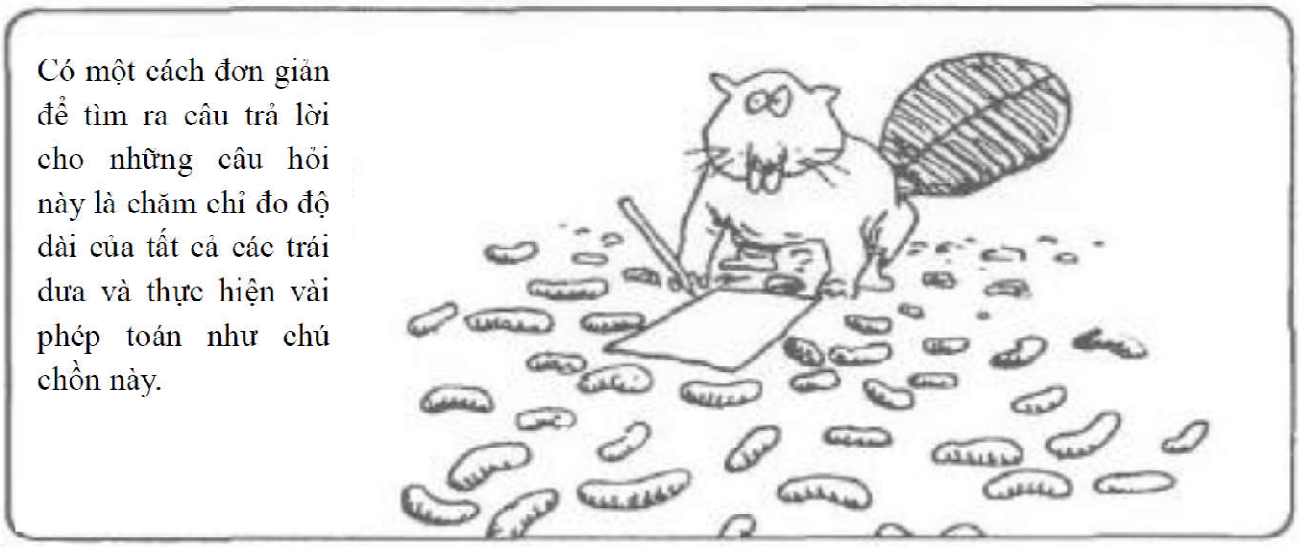
CHƯƠNG 6: CHỌN MẪU

Giờ đây, sau khi nghiên cứu kỹ các lý thuyết thông qua những ví dụ về đồng xu, xúc sắc, và những ý tưởng trừu tượng, có thể bạn sẽ ngạc nhiên vì tất cả các công cụ thống kê mà chúng tôi đã xây dựng đều được ứng dụng trong thực tế. Bây giờ chúng ta sẽ cùng nhau khám phá điều này…

Trong chương này, chúng ta sẽ bắt đầu tìm hiểu về tính ứng dụng của thống kê, điều này sẽ giúp bạn tiết kiệm được thời gian cũng như chi phí. Con người ghét việc phải lãng phí thời gian để làm những việc không cần thiết, và điều mà thống kê có thể làm là nói cho chúng ta biết chính xác là chúng ta có thể lười như thế nào.

Vấn đề ở chỗ thế giới là một tập hợp lớn các thông tin, khó có thể tìm được đúng thông tin mà chúng ta muốn.





Nhưng chúng ta không phải là những chú chồn - Chúng ta là những nhà thống kê! Chúng ta đang khéo léo tìm ra cách giải quyết…

Phương pháp của chúng tôi là **chọn mẫu**... một tập hợp con từ tổng thể, cách mà những người thăm dò ý kiến vẫn làm trong thời gian bầu cử.

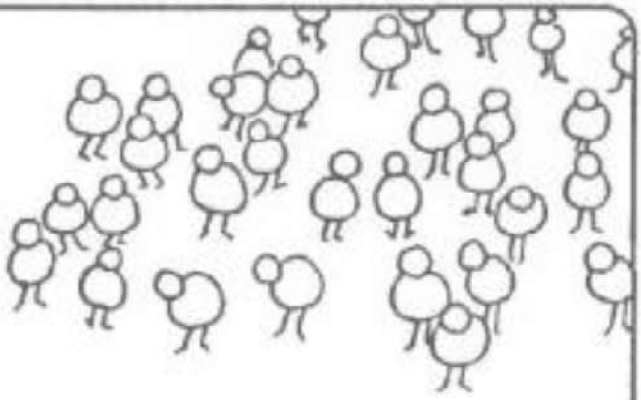
Một câu hỏi hiển nhiên là: Để các kết quả có ý nghĩa thì mẫu phải lớn cỡ nào?

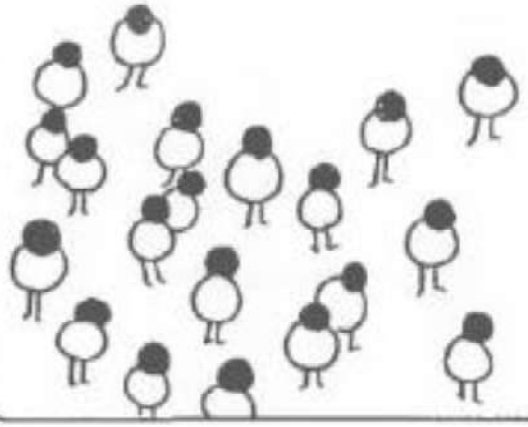
Và câu trả lời bạn nên khắc sâu vào bộ não là: Nếu n là số các quan sát trong mẫu, thì mọi thứ bị ảnh hưởng bởi:

$$\frac{1}{\sqrt{n}}.$$

THIẾT KẾ MẪU

Trước khi quan tâm đến số lượng, chúng ta nên chú ý tới chất lượng của mẫu cũng như là kích thước của nó. Làm thế nào để chúng ta có thể quả quyết với bản thân rằng chúng ta đang chọn được một mẫu đại diện?



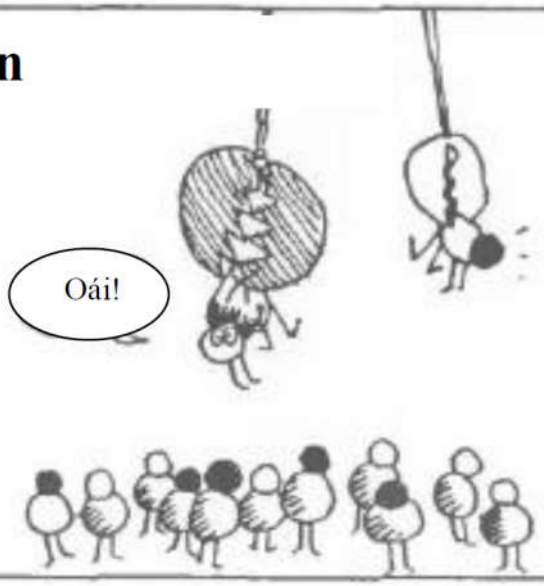


Bản thân quá trình chọn mẫu mang tính chất quyết định. Ví dụ, một cuộc điều tra về số người bỏ phiếu mà loại trừ người da đen thì sẽ vô tác dụng, và có nhiều cách làm hỏng, hoặc làm chệch mẫu.

Không để các bạn hồi hộp nữa, cách thức để giúp cho các kết quả thống kê đáng tin cậy chính là chọn mẫu một cách **ngẫu nhiên.**

MẪU ngẫu nhiên giản đơn

Giả sử chúng ta có một tổng thể lớn của các đối tượng và một quy trình chọn n đối tượng. Nếu quy trình đảm bảo tất cả các mẫu có thể của n đối tượng là công bằng như nhau, thì chúng ta gọi quá trình đó là quá trình chọn một **mẫu ngẫu nhiên giản đơn.**



Mẫu ngẫu nhiên giản đơn có hai đặc tính làm cho chúng tương phản với tiêu chuẩn mà chúng ta đo lường bằng các phương pháp khác:

1) Tính không chệch: Mỗi đơn vị có cơ hội được chọn như nhau.

2) Tính độc lập: sự lựa chọn một đơn vị không ảnh hưởng tới sự lựa chọn các đơn vị khác.

Thật không may, trong thế giới thực, mẫu độc lập, hoàn toàn không chệch rất khó tìm. Ví dụ, mẫu của việc điều tra những người bầu cử qua điện thoại là một mẫu bị chệch: Nó không tính đến những người bầu cử không dùng điện thoại và mẫu bị trùng khi có những người dùng nhiều hơn một số điện thoại.

Theo lý thuyết có thể chọn được một mẫu ngẫu nhiên bằng cách xây dựng **một dàn mẫu**: danh sách của mỗi đơn vị trong tổng thể. Chúng ta có thể chọn n đối tượng từ dàn mẫu một cách ngẫu nhiên.

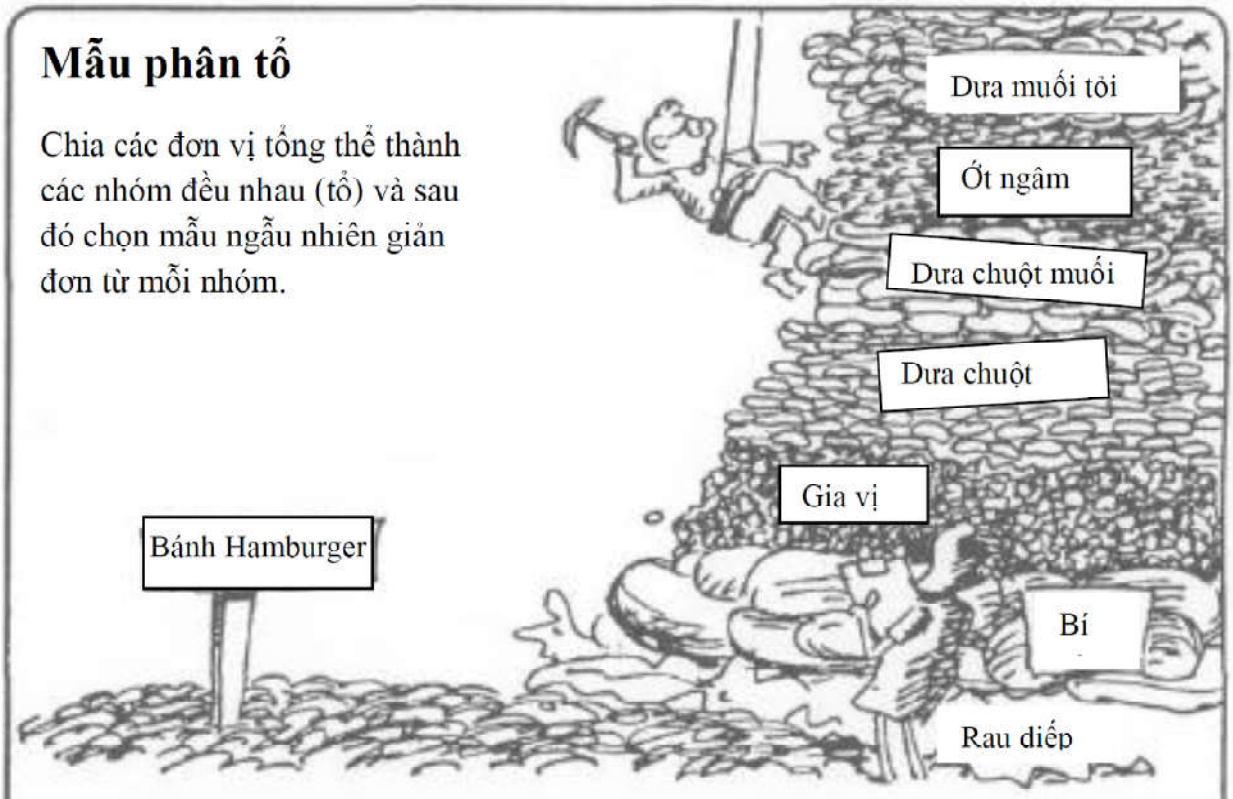
Tương tự, chúng ta có thể viết tất cả những cái tên lên trên những tấm các, đặt vào thùng và ngẫu nhiên rút ra n tấm.

Nhưng điều này không phải luôn luôn dễ dàng, làm một dàn mẫu có thể không loại trừ chi phí tốn kém, gây ra tranh luận, hoặc cũng có thể không khả thi. Ví dụ, nghiên cứu E.P.A về chất lượng nước cần một dàn mẫu của các hồ ở Mỹ, bởi vậy sau đó người ta phải giải quyết vấn đề:

Liệu có các cách khác để chọn mẫu **hiệu quả** và **lợi ích** hơn là chọn mẫu ngẫu nhiên giản đơn không? Vâng - Nếu câu trả lời là có thì bạn cũng đã biết được thêm vài điều về tổng thể đấy. Về trường hợp…

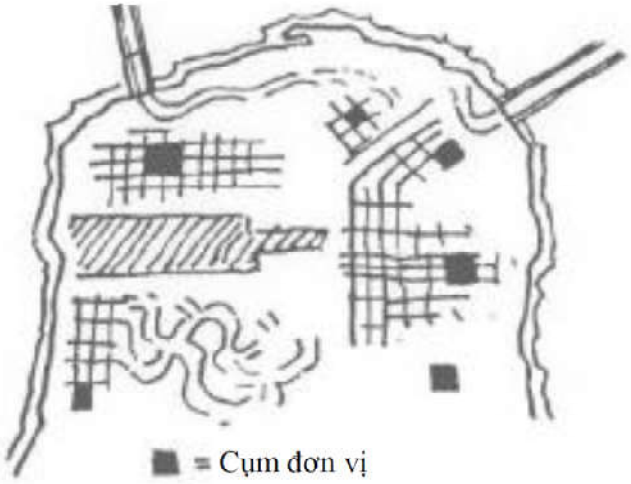
Mẫu phân tổ

Chia các đơn vị tổng thể thành các nhóm đều nhau (tổ) và sau đó chọn mẫu ngẫu nhiên giản đơn từ mỗi nhóm.



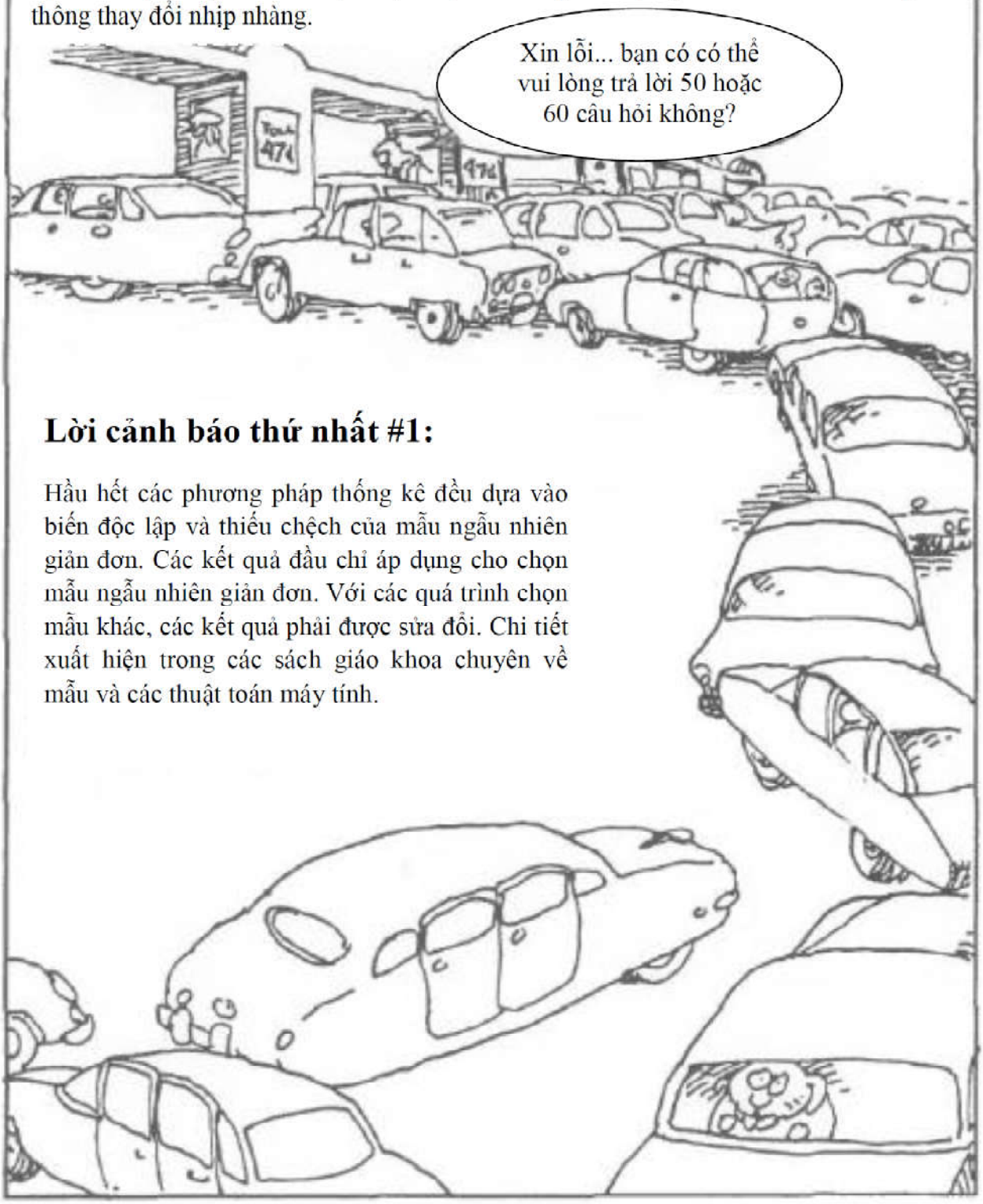
Ví dụ, tổng thể dưa góp có thể được phân tổ theo loại dưa. Trong mỗi loại hoặc mỗi tổ, kích thước trái dưa nên ít biến thiên.

Mẫu cụm (mẫu chùm) (*là cách chọn mẫu trong đó số đơn vị mẫu điều tra không phải là từng đơn vị riêng lẻ mà là từng cụm đơn vị*) phân loại tổng thể thành các cụm nhỏ, chọn mẫu ngẫu nhiên giản đơn từ mẫu cụm, và quan sát mọi thứ trong các mẫu cụm. Điều này có thể có lợi nếu chi phí di chuyển giữa các đơn vị mẫu ngẫu nhiên cao.



Ví dụ một cuộc điều tra về nhà ở thành phố, cuộc điều tra chia thành phố thành các khối, chọn mẫu ngẫu nhiên các khối, và quan sát tất cả mọi đơn vị nhà ở trong mỗi khối mẫu.

Mẫu hệ thống bắt đầu với một đơn vị được lựa chọn ngẫu nhiên, cứ sau k đơn vị tiếp theo lại chọn được một đơn vị nữa. Ví dụ, nghiên cứu giao thông trên đường cao tốc bằng việc kiểm tra hàng trăm lượt ô tô lưu thông qua một trạm thu phí đường bộ. Kế hoạch này có thể được thực hiện dễ dàng và hiệu quả hơn nếu cấu trúc giao thông thay đổi nhịp nhàng.



Lời cảnh báo thứ nhất #1:

Hầu hết các phương pháp thống kê đều dựa vào biến độc lập và thiếu chệch của mẫu ngẫu nhiên giản đơn. Các kết quả đầu chỉ áp dụng cho chọn mẫu ngẫu nhiên giản đơn. Với các quá trình chọn mẫu khác, các kết quả phải được sửa đổi. Chi tiết xuất hiện trong các sách giáo khoa chuyên về mẫu và các thuật toán máy tính.

Lời cảnh báo thứ hai #2:

Trong thiết kế ngẫu nhiên, có thể không có các phân tích mang tính thống kê độc lập, bất kể nó được thay đổi như thế nào. Vẻ đẹp của mẫu ngẫu nhiên là nó "đảm bảo tính thống kê" độ chính xác của cuộc điều tra.

Phương pháp sử dụng thông thường có xu hướng chệch: Nó được gọi là **mẫu cơ hội**. Tránh mọi phiền phức của việc thiết kế quá trình thực hiện người chọn mẫu cơ hội chỉ gom n đơn vị tổng thể đầu tiên.

Một ví dụ cổ điển về cuốn sách của Shere Hite: phụ nữ và tình yêu. 100000 bảng hỏi đã được chuyển tới các tổ chức phụ nữ (một mẫu cơ hội), chỉ 4.5% bảng hỏi được điền đầy đủ và thu hồi (trả lời chệch). Bởi vậy các "kết quả" được dựa trên một mẫu phụ nữ những người được khuyến khích để trả lời, cho dù với bất cứ lí do gì.

(Còn nữa)

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