CHƯƠNG 9: SO SÁNH HAI TỔNG THỂ

Nhưng điều gì gây thách thức cho các con số thống kê tương tự như sự phong phú trong ẩm thực. Giống như một chuyên gia ẩm thực, các nhà thống kê có thể "nếm" các thành phần theo cách của một vấn đề nào đó và tìm ra được cách hiệu quả nhất kết hợp chúng thành một công thức thống kê.

(Khi nói các hướng dẫn trong cuốn sách dạy nấu ăn cũng như các phương pháp thống kê đều hiệu quả có nghĩa là chúng đều đưa ra được giải pháp cho các tình huống khác nhau).

Trong chương này, chúng ta sẽ sử dụng phương pháp đối với thịt và khoai tây trên đây theo một vài công thức mới, điều này sẽ giúp chúng ta trả lời được các câu hỏi sau:

Liệu việc sử dụng thuốc aspirin đều đặn có giúp giảm nguy cơ bị đau tim hay không?

Việc thường xuyên dùng thuốc diệt cỏ có làm tăng diện tích cỏ trên mỗi hecta đồng ruộng không?

Và liệu rằng thu nhập giữa nam giới và nữ giới ở cùng một chức vụ có khác nhau không?

Đặc điểm chung của các câu hỏi này chính là: chúng đều có thể được trả lời thông qua việc so sánh hai mẫu ngẫu nhiên độc lập, mỗi mẫu được chọn ra từ một tổng thể khác nhau.

Có phun thuốc trừ sâu

Không phun thuốc trừ sâu

Còn ở phần cuối chương, chúng ta sẽ xem xét một cách so sánh hai trung bình khác, không liên quan tới hai mẫu ngẫu nhiên giản đơn…

So sánh TỶ LỆ THÀNH CÔNG

(hoặc tỷ lệ thất bại) của hai tổng thể.

Chúng ta bắt đầu với một thí nghiệm, một phần trong công việc nghiên cứu ở Harvard, mục đích nghiên cứu nhằm đánh giá hiệu quả của thuốc aspirin trong việc giảm thiểu bệnh đau tim. Giống như hầu hết các biểu hiện lâm sàng, nguy cơ một vài bệnh nhân bị mắc bệnh - trường hợp này là bệnh đau tim - thường là rất nhỏ trong một vài năm nghiên cứu. Nhưng chúng ta lại muốn trả lời câu hỏi trên một cách nhanh chóng! Vậy chúng ta phải làm gì?

Đơn giản, nhưng tốn kém, giải pháp chính là kiểm tra một số lượng lớn các cá nhân trong khoảng thời gian ngắn. Trong quá trình nghiên cứu này, 22071 đơn vị (toàn bộ các bác sỹ tình nguyện) được phân chia ngẫu nhiên thành hai nhóm.

Nhóm một nhận được thuốc giả - một loại thuốc có thành phần chính là aspirin nhưng thực thế không phải là aspirin.

Nhóm hai mỗi ngày được nhận một viên aspirin

Trung bình toàn bộ quá trình dài gần 5 năm*, những người khảo sát đã ghi chép lại các câu trả lời: bị đau tim hoặc không bị đau tim.

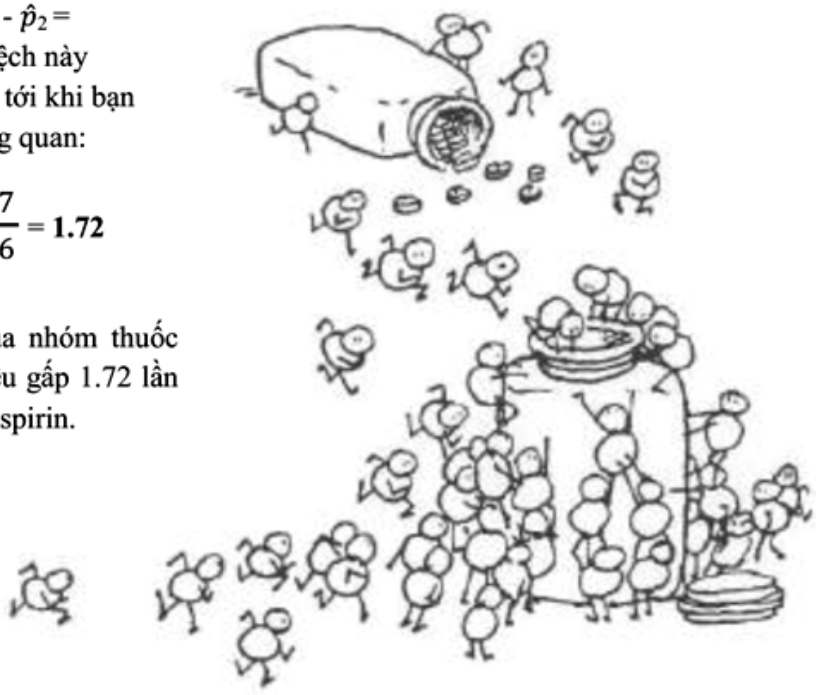
Kết luận: các con số dưới đây, chúng tôi đã kết hợp trong bảng cả các câu trả lời bị đau tim và không bị đau tim.

	Đau tim	Không đau tim	n	
Thuốc giả	239	10795	11034	$\hat{p}_1 = \frac{239}{11034} = 0.0217$
Aspirin	139	10898	11037	$\hat{p}_2 = \frac{139}{11037} = 0.0126$

Sự khác biệt quan sát được về tỷ lệ thành công là $\hat{p}_1 - \hat{p}_2 = 0.0091$. Sự chênh lệch này dường như nhỏ chỉ tới khi bạn để ý tới rủi ro tương quan:

$$\frac{\hat{p}_1}{\hat{p}_2} = \frac{0.0217}{0.0126} = \mathbf{1.72}$$

Các thành viên của nhóm thuốc giả bị đau tim nhiều gấp 1.72 lần nhóm dùng thuốc aspirin.



* Việc nghiên cứu đã bị dừng lại sớm hơn bởi kết quả tích cực này. Nhưng đó là điều không hề khôn ngoan và thiếu tính thực tế khi phủ nhận các kết quả của nhóm dùng thuốc giả định.

Mô hình: Các quan sát của nhóm thuốc giả và nhóm dùng aspirin là những mẫu độc lập từ hai tổng thể nhị thức. Để nhất quán, chúng ta coi bị đau tim là một **thành công** (!)

Khả năng thành công của nhóm dùng thuốc giả = p_1

Khả năng thành công của nhóm dùng thuốc aspirin = p_2

Mục tiêu là ước lượng được chính xác sự khác nhau của $p_1 - p_2$.

Với mỗi tổng thể (thực tế là với các mẫu lớn của tổng thể chung), chúng ta có các biến ngẫu nhiên đặc trưng:

X_1 Số thành công trong tổng thể thứ nhất

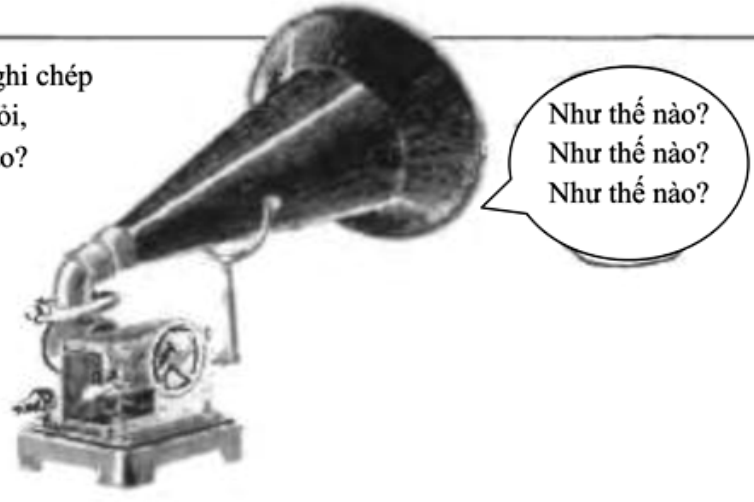
X_2 Số thành công trong tổng thể thứ hai

$\hat{P}_1 = \frac{X_1}{n_1}$ Tỷ lệ thành công trong tổng thể thứ nhất

$\hat{P}_2 = \frac{X_2}{n_2}$ Tỷ lệ thành công trong tổng thể thứ hai

Và ước lượng tỷ lệ sai khác là: $\hat{P}_1 - \hat{P}_2$

Còn bây giờ, coi như bản ghi chép trên bị hỏng, chúng ta tự hỏi, $\hat{P}_1 - \hat{P}_2$ phân bố như thế nào?



Phân bố mẫu của $\hat{P}_1 - \hat{P}_2$

Với nhiều mẫu lớn thì $\hat{P}_1 - \hat{P}_2$ xấp xỉ phân bố chuẩn hơn là trường hợp chỉ có duy nhất một mẫu. Thông thường, chúng ta có thể tính z-biến đổi để tìm biến ngẫu nhiên chuẩn tắc (xấp xỉ).

$$z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sigma(\hat{P}_1 - \hat{P}_2)}$$

Nhưng làm thế nào để tìm được độ lệch chuẩn trong mẫu số trên?

Vì hai mẫu độc lập với nhau, bởi vậy hai biến ngẫu nhiên $\hat{P}_1$, $\hat{P}_2$ và hai phương sai cũng độc lập với nhau:

$$\sigma^2(\hat{P}_1 - \hat{P}_2) = \sigma^2(\hat{P}_1) + \sigma^2(\hat{P}_2)$$

Nên

$$\sigma(\hat{P}_1 - \hat{P}_2) = \sqrt{\sigma^2(\hat{P}_1) + \sigma^2(\hat{P}_2)}$$

Và bây giờ, hiểu được phân bố của kiểm định thống kê, chúng ta có thể tiến hành ước lượng các khoảng tin cậy và thực hiện kiểm định các giả thuyết cho rằng thuốc aspirin có thể giúp giảm chứng bệnh đau tim.

Còn nữa)

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