



# WFSJ

World Federation of Science Journalists

## Khóa học trực tuyến về Báo chí khoa học

Thực hiện bởi WFSJ và SciDev.Net



## Bài 9

### Hiểu biết về dữ liệu thống kê

Nguyên bản: Hans Van Maanen

Chỉnh sửa và Việt hóa: TS. Nguyễn Đức An



Time for  
**E-LEARNING**



<b>9.1</b>	Giới thiệu .....	3
<b>9.2</b>	Các con số .....	4
<b>9.3</b>	Tỉ lệ phần trăm .....	5
<b>9.4</b>	Các giá trị trung bình .....	6
<b>9.5</b>	Độ phân tán trong dãy số .....	7
<b>9.6</b>	Tỷ lệ: nguy cơ tương đối .....	8
<b>9.7</b>	Vận may .....	9
<b>9.8</b>	Sự tương quan so với quan hệ nhân quả .....	10
<b>9.9</b>	Ý nghĩa thống kê .....	11
<b>9.10</b>	Vài phạm bẫy cuối cùng .....	12
<b>9.11</b>	Kết luận .....	13
<b>9.12</b>	Câu hỏi tự luận (1-6) .....	14
<b>9.13</b>	Đáp án câu hỏi tự luận (1-6) .....	15
<b>9.14</b>	Câu hỏi tự luận (7-12) .....	16
<b>9.15</b>	Đáp án câu hỏi tự luận (7-12) .....	17
<b>9.16</b>	Bài thực hành .....	18



## 9.1 Giới thiệu

Nhà báo, nhất là nhà báo khoa học, có thực sự cần biết ít nhiều về thống kê không? Không, nếu họ có một người bạn là nhà thống kê, người họ có thể gọi bất cứ lúc nào dù là ngày hay đêm. Những người khác đều phải hiểu ít nhất những nguyên tắc cơ bản về tỷ lệ phần trăm, các chỉ số trung bình, độ lệch và kiểm định ý nghĩa thống kê. Có thể là họ không cần sử dụng thống kê một cách chủ động – không ai kỳ vọng nhà báo khoa học làm các kiểm định chi bình phương (chi-square) cho dữ liệu – nhưng tri thức thụ động sẽ luôn luôn có ích.

Bài này sẽ cố gắng cung cấp cho bạn những nguyên tắc sơ đẳng về các con số và dữ liệu thống kê, nhưng nó không phải là “thống kê dành cho người rơm” theo cách thông thường. Vấn đề này quá rộng để nói hết trong một chương, và tư liệu liên quan đến thống kê cơ bản trên Wikipedia là khá rõ ràng và đáng tin cậy. Thay vào đó, bạn sẽ tìm hiểu về những yếu tố có thể gây sai sót – và thực tế là rất thường gây sai sót – khi nhà báo và cả các nhà khoa học tường thuật các vấn đề/hiện tượng khoa học.

“

Xong bài này, bạn sẽ cảnh giác hơn với tất cả các tuyên bố dựa trên số liệu trong khoa học, và bạn sẽ biết đặt những câu hỏi khó hơn. Ít nhất, bạn sẽ không hoàn toàn thiếu khả năng kiểm soát khi các số liệu xuất hiện. Và chúng rất thường xuất hiện.

”





### Kiểm định các con số.

Hãy nhớ rằng nhiều con số lớn có thể đã được tính nhầm ngay tại hiện trường, chứ chẳng phải đo lường chính xác. Số người tham gia tại một cuộc biểu tình, số bé gái bị bắt cóc mỗi năm, lượng khán giả truyền hình tại Thế vận hội Olympic...

Khi gặp những con số như trên, một câu hỏi luôn cần thiết là: làm thế nào họ biết? Điều đó có thể biết được không? Và con số đó là lớn hay không trong bối cảnh này?



### 9.2 Các con số

Kích thước hạt bụi là bao nhiêu? Có bao nhiêu tế bào máu hoạt động trong cơ thể người? Diện tích một sân bóng đá là bao nhiêu? Độ cao mặt trăng là bao nhiêu?

**Đưa con số vào bối cảnh.** Một trong những kỹ năng quan trọng nhất đối với nhà báo khoa học – cũng như với nhà báo trong bất cứ lĩnh vực nào – là nhanh chóng đưa con số vào trong bối cảnh. Đó là cách duy nhất để nắm bắt chúng, để hiểu được tầm vóc sự vật và hiện tượng liên quan, và dựng nên ý nghĩa cho chúng.

Các con số là những khái niệm trừu tượng (vì trừu tượng nên chúng hay được sử dụng một cách dễ dàng), nhưng để chúng có ý nghĩa đối với độc giả, bạn sẽ phải mang chúng gắn với thực tại đời sống hơn. Như chúng ta thường nghe: **hãy thử hỏi người vào con số.**

Sẽ rất khó khăn để hiểu kích thước một khu vườn rộng 0,735 ha là thế nào, nhưng mọi người sẽ hình dung được khu vườn đó ra sao nếu bạn nói nó có kích thước bằng một sân bóng đá. Một hạt bụi có kích thước khoảng 0,125mm (tức là một phần tám mm). Vì vậy, trên diện tích một mm<sup>2</sup> (nhỏ hơn dấu chấm câu), chúng ta có thể tính được có  $8 \times 8 = 64$  hạt bụi.

Một cách luôn hữu ích cho bất kỳ một câu chuyện nào là “tái nhân cách hóa” các con số bằng một so sánh hợp lý. Nó cũng rất quan trọng vì nó cho bạn một cách để dễ dàng kiểm tra độ lớn các con số. Chỉ bằng cách này, bạn mới nhận biết được một con số – chẳng hạn như nói 1000 lần – là quá lớn hay quá nhỏ. Một ngàn lần, thông thường, là rất nhiều. Chúng ta có thể dễ dàng sống trong một triệu giây – nó ít hơn 12 ngày – nhưng một tỷ giây là 32 năm và một nghìn tỷ giây là 32.000 năm. 32.000 năm trước, người Neanderthal cuối cùng qua đời.



### 9.3 Tỷ lệ phần trăm

Tỷ lệ phần trăm chẳng phải đơn giản như khi nhìn thoạt qua. 1% là một phần của một trăm, cho nên 8% của 80 kg là 6,4 kg. Và nếu chúng ta tăng lượng bò nuôi từ 50 lên 70, chúng ta đã tăng lên  $(70 - 50)/50 = 0.4$ , tương đương 40%.

Điều đầu tiên cần hiểu là không phải ai cũng hiểu 40% nghĩa là gì. Ở Đức, một nghiên cứu cho thấy, hóa ra chỉ có hơn nửa dân số (54%) hiểu con số này là gì. **Nhiều thứ đơn giản với chúng ta có thể không đơn giản với độc giả.**

Và tỷ lệ phần trăm hiếm khi đơn giản như trên. Chúng ta chia tử số cho mẫu số, rồi thể hiện kết quả qua một con số khác. Vì thế, một số thông tin ban đầu đã thất lạc đâu đó trong quá trình, đặc biệt là khi từ ngữ diễn đạt nó không chính xác – và đây là một điều thường xảy ra.

Một nhà khoa học nói với chúng ta rằng thuốc trị trầm cảm dẫn tới nhiều sự cố tình dục (như bất lực hay mất cảm hứng tình dục) trong 40% ca bệnh. Nhưng nói như thế có nghĩa là gì? Số sự cố tình dục ở một người dùng thuốc sẽ cao hơn 40% so với một người không dùng thuốc, hay 40% số người sử dụng thuốc sẽ gặp phải những sự cố này?

**Bối cảnh.** Giống như các con số, việc đầu tiên chúng ta cần làm là đưa tỷ lệ phần trăm vào bối cảnh. Tăng 8% có thể có nhiều ý nghĩa nếu đó là trọng lượng cơ thể. Nhưng tăng 40% có thể là một biến động không đáng kể nếu đó là huyết áp hay hàm lượng cholesterol.

Thậm chí, nếu nguy cơ đột quỵ tăng 40%, nó cũng sẽ không mang ý nghĩa quan trọng, vì bản thân nguy cơ xảy ra đột quỵ là rất thấp. 40% nghe có vẻ ấn tượng, nhưng điều đó không có nghĩa là nó đáng quan tâm, hay như người ta hay nói, có giá trị về mặt lâm sàng.

Thế nên, không nên hoảng hốt bán nhà phố để chuyển sang nơi yên tĩnh hơn, khi đọc những bản tin như “Tiếng ồn làm tăng nguy cơ đột quỵ” trên *Dân Trí*, được dịch từ tờ *Daily Mail*:  
[<http://dantri.com.vn/suc-khoe/tieng-on-lam-tang-nguy-co-dot-quy-303220.htm>].

**Điểm phần trăm.** Đôi khi “phần trăm” (percent) và “điểm phần trăm” (percent point) bị lẫn lộn, vô tình hay cố ý. Nếu mức thất nghiệp năm ngoái là 4,8% và năm nay là 6%, thì tỉ lệ tăng sẽ là  $(6-4,8)/4,8 = 0.25$ , tương đương 25%. Hay nó chỉ tăng  $6 - 4.8 = 1.2\%$ ? Chúng ta có thể sử dụng cả hai con số này, nhưng để tránh nhầm lẫn, các nhà kinh tế sử dụng thuật ngữ “điểm phần trăm” trong trường hợp sau.

Ở đây, một chính phủ đương quyền sẽ sử dụng 1.2 điểm phần trăm (vì nó nghe có vẻ ít hơn nhiều so với 25 phần trăm), trong khi phe đối lập thì sẽ làm ngược lại! Không có quy tắc nào cả – ngoại trừ việc yêu cầu nguồn tin cung cấp dữ liệu đằng sau những con số đó. Hãy nhớ là các chính quyền có thể thay đổi định nghĩa về thất nghiệp và lạm phát để chúng trông có vẻ thỏa mãn hơn, hay ít tai hại hơn.



Đôi khi, nhà báo cần biết đặt những câu hỏi sắc cạnh để có thể nhìn sâu hơn ở phía sau mặt con số.





Các giá trị trung bình có xu hướng làm lu mờ sự biến thiên tự nhiên trong tập hợp số liệu. Đôi khi nó tốt, nhưng đôi khi nó làm mất đi quá nhiều thông tin – độ biến thiên có thể thú vị hơn và quan trọng hơn so với giá trị trung bình. Tất nhiên, các nhà khoa học đã phát minh ra một cách để khắc phục hạn chế này: họ sử dụng một chỉ số thống kê khác, gọi là độ lệch, để tóm tắt sự thay đổi này.



### 9.4 Các giá trị trung bình

Khi nói đến các giá trị “trung bình” (average trong tiếng Anh), người ta thường hiểu ngầm chúng là cùng cách tính như nhau. Trên thực tế, có đến vài loại “giá trị trung bình” khác nhau – ở đây chúng ta sẽ làm quen với hai loại giá trị trung bình phổ biến hơn: giá trị bình quân hay trung bình cộng (gọi là mean trong tiếng Anh) và giá trị trung vị hay trung bình giữa (median).

**Giá trị bình quân.** Một cậu bé có 4 xu, cậu thứ hai có 5 xu và cậu thứ ba có 9 xu. Bình quân, mỗi cậu có  $(4+5+9)/3 = 6$  xu. Đó là giá trị bình quân (hay còn gọi là trung bình cộng). Bình quân, mỗi cậu bé có 6 xu, mặc dù không cậu nào thực sự có sáu xu.

Cũng như tỉ lệ phần trăm, các giá trị bình quân trông rất đơn giản, nhưng chúng có thể “đánh lừa” chúng ta. Trong ví dụ trên, nếu cậu bé thứ ba được cho gấp đôi số tiền hiện có, số tiền bình quân sẽ tăng từ 6 lên  $(4+5+18)/3 = 9$  xu. Hai trong ba cậu bé không có thêm xu nào, nhưng vẫn có “số tiền bình quân” cao hơn trước. Có gì đó không ổn trong con số 9 xu này chẳng?

Nếu 19 người lao động mỗi người kiếm được 100 đô la và ông chủ được 2100, thì thu nhập bình quân trong công ty sẽ là  $(19 \times 100 + 2100)/20 = 200$  đô la. Dĩ nhiên, đó không phải là cách công đoàn nhìn vấn đề, bởi rõ ràng là 95% người trong công ty chỉ kiếm được nửa mức bình quân đó. Mức bình quân trong trường hợp này là vô nghĩa vì nó bị đẩy lên quá cao so với thực tế chỉ vì mức thu nhập cực kỳ cao của chỉ một cá nhân.

Năm 2007, một lãnh đạo tỉnh Trà Vinh vui mừng thông báo cho một đám đông dân chúng là thu nhập bình quân đầu người trong tỉnh đã đạt được 800 đô la/năm. Bản thân con số 800 đô đó đã không nhiều rồi, nhưng còn một vấn đề chua hơn, khi một ông nông dân đứng lên nói nôm na: “Một người ăn nguyên một con gà, một người chỉ đứng nhìn, tính bình quân mỗi người ăn được... nửa con gà. Nghe 800 USD tui ham lắm, nhưng chùng nào tui và gia đình tui mới có được?”

Gần như ai trong chúng ta cũng đều có hai chân. Chỉ có một thiểu số không may sẽ có ít hơn hai chân. Nhưng tính kiểu bình quân thì rõ ràng số chân bình quân mỗi người chắc chắn không thể là hai!

**Giá trị trung vị.** Trong những trường hợp trên, chúng ta sẽ phải sử dụng giá trị trung vị (hay còn gọi là trung bình giữa). Nói nôm na: “trung vị” là đứng giữa: một nửa trường hợp sẽ thấp hơn hoặc bằng nó, và một nửa sẽ cao hơn hoặc bằng. Đối với các cậu bé, chúng ta có giá trị trung vị là 5, cả trước và sau khi cậu bé thứ ba được tăng gấp đôi số tiền. Trong trường hợp công ty, giá trị trung vị là 100 đô la. Trong trường hợp chân người, con số đó sẽ là hai.

Giá trị trung vị trong những trường hợp này phản ánh đúng thực tế hơn vì chúng không nhạy cảm với các giá trị cực lớn hoặc cực nhỏ trong dãy số. Cho dù thu nhập ông chủ là 2100 hay 21.000 đô la, nó cũng sẽ đứng ở cuối dãy số. Cũng vậy, nếu cậu bé thứ ba có đến 90 xu hay nhiều hơn.

**Nhà báo với các giá trị “trung bình”.** Khi tường thuật tin tức cho công chúng, nhà báo sẽ hiếm khi sử dụng những thuật ngữ chuẩn xác như “trung vị” hay “trung bình cộng”. Thay vào đó ta thường dùng thuật ngữ chung chung và gần gũi với người dân hơn là “trung bình” – thu nhập trung bình, chỉ tiêu trung bình, độ tuổi trung bình, chiều cao trung bình... Nó giống như chữ average chung chung trong tiếng Anh.

Nhưng khi xử lý tin bài đăng sau mặt báo, bạn hãy giữ cái đầu tỉnh táo để phân định mỗi con số trung bình đó chính xác là cái gì. Thiếu hiểu biết những điều trên, ta sẽ dễ bị đánh lừa khi xử lý các “giá trị trung bình”.

Nhân viên cứu hỏa phàn nàn rằng họ bị giảm đáng kể về “thu nhập trung bình” vào năm ngoái, và yêu cầu tăng lương. Nhưng sau đó, ai đó đã phát hiện ra rằng năm ngoái rất nhiều lính cứu hỏa mới và trẻ được thuê – và những người này bắt đầu với mức lương thấp, khiến cho “thu nhập trung bình” (tức là “thu nhập bình quân” trong trường hợp này) bị đẩy xuống.

Ở nhiều nước phương Tây, độ tuổi trung bình của các bà mẹ sinh con đầu lòng đang tăng: “Số bà mẹ cao tuổi hơn đang tăng”, các tit báo đại loại như vậy xuất hiện. Nhưng độ tuổi trung bình tăng vì phụ nữ có con đầu lòng muộn hơn, hay vì có ít các bà mẹ trẻ hơn? Hay cả hai? Các nhà khoa học so sánh những năm nào với nhau, và lý do tại sao lại chọn chính xác hai năm này?



## 9.5 Độ phân tán trong dãy số

Các cậu bé có 4,5 và 9 xu, bình quân là 6 xu. Các cô bé có 5,6 và 7 xu, cũng có bình quân là 6 xu. Nhưng sự chênh lệch trong phân bố số tiền xu giữa các cậu bé lớn hơn nhiều – tức là, có nhiều biến thiên về “tài sản” giữa các cậu bé hơn giữa các cô bé.

Nhiệt độ bình quân ở hai thành phố đều là 25 độ, nhưng thời tiết hai nơi có thể hoàn toàn khác nhau. Một nơi, nhiệt độ quanh năm chủ yếu ở trong mức 20-30 độ (không có mùa đông và mùa hạ tương đối mát mẻ). Ở nơi kia, nó là 10-40 độ (mùa đông rõ rệt và mùa hạ nóng bức). Đó là do sự cách biệt về mức phân tán trong hai dãy dữ liệu thô về nhiệt độ.

**“Quy tắc bỏ túi” về độ lệch chuẩn.** Các nhà thống kê đưa ra một chỉ số gọi là “độ lệch chuẩn” (standard deviation) để mô tả đặc tính này trong sự phân bố số liệu thô (tức số liệu thu thập được từ từng cá thể/trường hợp liên quan, chưa qua phân tích thống kê). Trong hầu hết các bài báo khoa học, chúng ta có những bảng dữ liệu dài với các chỉ số thống kê như: “Tuổi bình quân: 65, độ lệch chuẩn: 5 năm”.

Điều quan trọng nhất với các nhà báo là hiểu rằng **hầu hết (95%) phần tử trong tập hợp số liệu phân bố trên phạm vi có độ rộng bằng 4 lần độ lệch chuẩn.** Một nửa trong 95% đó nằm trong hai độ lệch chuẩn dưới giá trị bình quân, và nửa còn lại nằm trong hai độ lệch chuẩn phía trên giá trị bình quân. Trong trường hợp trên, phạm vi đó là  $4 \times 5 = 20$  năm. Nghĩa là, 95% những người tham gia nghiên cứu này ở độ tuổi từ  $65 - 2 \times 5 = 55$  đến  $65 + 2 \times 5 = 75$  tuổi.

Có thể xem đó như một quy tắc bỏ túi – dựa trên kinh nghiệm, thô sơ nhưng tiện lợi – trong việc đọc dữ liệu thống kê. Nếu bạn muốn biết chính xác hơn thì mô hình phân bố phần tử tập hợp dữ liệu thường sẽ như sau:

- Khoảng 95% phần tử tổng thể sẽ nằm trong khoảng bao phủ bởi 4 độ lệch chuẩn.
- 99 phần trăm nằm trong khoảng 6 độ lệch chuẩn.
- 68% nằm trong hai độ lệch chuẩn.

Vì vậy, chỉ với một chỉ số nhỏ gọn là độ lệch chuẩn, cho chúng ta biết rất nhiều về thực tế.

Đôi khi thay vì “độ lệch chuẩn”, người ta sử dụng một chỉ số thống kê khác gọi là phương sai: phương sai đơn giản chỉ là bình phương của độ lệch chuẩn.

**Chỉ số z.** Tất nhiên, có một số cạm bẫy khiến việc đọc hiểu độ lệch chuẩn theo quy tắc bỏ túi ở trên một cách mù quáng. Nhưng nếu xử lý thận trọng, chúng lại là những công cụ tuyệt vời.

Nếu bạn tính một sự thay đổi nào đó theo tỉ lệ phần trăm và chia nó cho độ lệch chuẩn, bạn sẽ có được một ấn tượng đầu tiên rất tốt về ý nghĩa của kết quả. Chỉ số IQ trung bình ở một quốc gia là 100 điểm với độ lệch chuẩn là 15 điểm. Vài năm sau, điểm IQ tăng lên 105, tức là tăng 5%. Thay đổi này tương đương với  $5/15 = 0,33$  độ lệch chuẩn. Theo thuật ngữ thống kê, con số 0.33 này được là chỉ số z.

Trong khoa học xã hội, hệ số z ở mức 0.2 thường được coi là “hiệu ứng yếu”, 0.5 được coi là “trung bình” và 0.8 là một hiệu ứng “mạnh”. Trong ví dụ trên, chỉ số z là 0.33, không phải quá thấp, nghĩa là mức độ tăng IQ ở quốc gia trong những năm đó không phải là không đáng kể.



Các nhà khoa học rất yêu tỷ lệ (rate) và tỷ số (ratio). Rủi ro tương đối, tỷ số nguy cơ, tỷ số rủi ro, tỷ lệ tử vong chuẩn... Một phóng viên “tội nghiệp” phải làm gì trước những khái niệm nghe như mơ hồ này?



### 9.6 Tỷ lệ: nguy cơ tương đối

Một lần nữa, cũng như với tỷ lệ phần trăm và các giá trị trung bình, hãy giữ một cái đầu lạnh, yêu cầu nguồn tin cung cấp các dữ liệu căn bản đầu tiên sau đó, và cố gắng đặt tất cả trong bối cảnh.

Một tỷ lệ phóng viên y học thường gặp là nguy cơ tương đối (relative risk). Hãy bắt đầu với một ví dụ thực tế. Uống statin làm giảm nguy cơ đột quỵ khoảng 50% trong vòng bốn năm – nguy cơ đột quỵ tương đối ở nhóm uống statin so với nhóm uống placebo (nhóm uống giả dược để so với nhóm được điều trị) là 0.5. Trong số những người uống placebo trong bốn năm, có khoảng 3% bị đột quỵ. Trong nhóm những người dùng statin, tỉ lệ này giảm xuống 1.5%.

Nói cách khác, 30 trong 1000 người thuộc nhóm giả dược bị đột quỵ, so với 15 trong số 1000 người dùng statin. Vì vậy, để ngăn ngừa 15 cơn đột quỵ, cần 1000 người dân dùng statin – như vậy, số lượng điều trị cần thiết để ngăn chặn một cơn đột quỵ (trong thời gian bốn năm) là  $1000/15 = 67$ . Cách tốt nhất để hình dung hiệu ứng này, là vẽ ra 67 người tí hon, và chọn một – đó là một người dùng statin. 66 người khác chỉ chịu tác dụng phụ của thuốc.

Ví dụ này cho thấy cần phải phân biệt rõ ràng giữa nguy cơ tương đối và nguy cơ tuyệt đối. Đôi khi phải tìm rất nhiều mới ra các dữ liệu đúng – và đôi khi, bạn không thể tìm thấy chúng ở bất cứ đâu. Nhiều bài báo khoa học – đặc biệt là về lợi ích và nguy hiểm của thực phẩm (như cà phê, trà, sữa, thịt đỏ) cũng như của truyền hình hay hoạt động thể chất – đưa ra những nguy cơ tương đối rất lớn, nhưng, sau rất nhiều tính toán, hóa ra mức giảm/tăng nguy cơ tuyệt đối không đáng kể.





### 9.7 Vận may

Nếu có một điều để phân biệt người làm thống kê với người bình thường, đó là vận may (luck). Nhà thống kê biết rằng vận may, tức sự xuất hiện ngẫu nhiên, đóng một vai trò lớn hơn nhiều trong cuộc sống chúng ta và trong tất cả các loại sự kiện mà chúng ta nghĩ là có thể.

Nếu ai đó trúng số hai lần liên tiếp, nó sẽ thành một tin nổi bật. Nhưng người làm thống kê sẽ lưu ý rằng, nếu tính gộp tất cả các loại xổ số đang diễn ra trên thế giới, đây là điều đôi khi sẽ xảy ra, ở nơi nào đó. Chúng ta gặp một người nào đó mà ta thấy trong mơ, ta thấy đó quả là bất ngờ lớn và thế giới thật nhỏ bé. Nhưng nhà thống kê sẽ nói mỗi đêm chúng ta mơ thấy nhiều người và, vì vậy, gặp một trong số họ trong thực tế cũng không phải là một trùng hợp gì to tát cho lắm.

Hơn nữa, mọi người luôn luôn có xu hướng nhìn thấy nhiều sự kiện như thể chúng có hệ thống, ngay cả khi chúng không theo hệ thống gì cả (tức là do tình cờ). Đó là cách bộ não chúng ta hoạt động – ta nhìn thấy các gương mặt trên mặt trăng, những con hổ trong bụi cây, và bất cứ điều gì đó xảy ra ba lần liên tiếp là "đáng chú ý". Nhiều điều mê tín dị đoan truyền miệng từ người này qua người khác ở Việt Nam cũng do vậy.

Điều đó đặc biệt rất dễ xảy ra khi người ta tìm cách diễn giải một quan hệ nhân quả. Nhiều trường hợp bệnh bạch cầu được phát hiện gần một nhà máy điện và bạn sẽ gặp các tí báo đại loại như: "Nhà máy điện gây bệnh bạch cầu". Đội bóng đá có huấn luyện viên mới và ngay lập tức họ giành chiến thắng ba trận liên tiếp – nhờ gì vậy, huấn luyện viên hay cơ may thuần túy?

Trong một cuộc đánh cược, ngay cả khi một đồng xu được thả lên xuống một cách vô tư, nó vẫn có thể lật ngửa bảy lần liên tiếp – xác suất xảy ra việc này chẳng phải là quá thấp, nhưng hãy thử giải thích điều đó với một đám đông đang nghi ngờ một điều gì mờ ám đằng sau đồng xu mà xem! **Người ta không chỉ đánh giá thấp quyền năng của cơ may mà còn không có cách bẩm sinh nào để đối phó với nó.**

Điều tương tự cũng xảy ra với nguy cơ. Người ta có xu hướng coi nhẹ những nguy cơ mà họ đã quen thuộc và có khả năng kiểm soát trong chừng mực nào đó (như lái xe). Họ lại thổi phồng những nguy cơ mà họ không quen thuộc và buộc phải chấp nhận (như ô nhiễm môi trường).

Cũng như khi thể hiện tỷ lệ phần trăm và giá trị trung bình, khi tường thuật về vận may hay nguy cơ, bạn nên tìm cách chuyển tải chúng sang một thứ ngôn ngữ mà độc giả đã quen thuộc. Đừng nói nguy cơ bệnh này là "1 trong 20". Ít ra hãy nói: "Trong một đám đông trăm người, sẽ có năm người bị bệnh". Đừng nói "1 trong 10.000". Hãy so sánh nó với số thương vong do xe hơi, thương vong do sét đánh...

Đôi khi, có những con số về nguy cơ hay thực trạng làm con người ta hoang mang, nhưng khi được đặt vào bối cảnh so sánh, nó chẳng đáng là gì bằng những con số tương đương khác.



Năm 2009, cả Việt Nam rúng động và hoảng sợ vì dịch cúm heo (A/H1N1), với 11.000 ca bệnh và 59 người tử vong. Nhưng, như Phan Kim Sơn ở báo *Sài Gòn Tiếp Thị*, phân tích: "Trong khi đó, bệnh lao ít khi gây tử vong nhanh, thế nhưng bệnh lại lây lan trong cộng đồng quanh năm suốt tháng. Ước tính hàng năm nước ta có 180.000 ca mắc lao mới và 32.000 người tử vong vì lao (gấp 540 lần số tử vong vì bệnh cúm A/H1N1!) nhưng không mấy người xem nguy cơ này là đáng sợ". Xem thêm ở <http://sgtt.vn/Khoa-giao/158558/Toi-oi-dung-so.html>.





## “ VÍ DỤ:

Các cột anten điện thoại di động khiến người ta sinh đẻ nhiều hơn? Dở hơi? Ấy vậy mà người ta tìm thấy rằng cứ một cột anten được dựng lên trong một khu vực thì bình quân có thêm 17.6 trẻ sơ sinh trong khu vực đó so với tỉ lệ sinh đẻ toàn nước Anh. Xem bài sau trên *The Guardian* [<http://www.guardian.co.uk/science/blog/2010/dec/17/mobile-phone-masts-birth-rate>], để hiểu vì sao mà các nhà báo đưa ra các kết luận sai từ các số liệu tương quan.



### 9.8 Sự tương quan so với quan hệ nhân quả

Sai lầm lớn và phổ biến nhất trong thống kê trên báo chí là đánh đồng sự tương quan (correlation) với quan hệ nhân quả (causation). Ở đâu có ách tắc giao thông là có cảnh sát, nhưng sẽ là phi lý nực cười nếu dựa trên tương quan đó để nói là cảnh sát nói chung là nguyên nhân gây ách tắc.

Tìm những ví dụ ngớ ngẩn tương tự chẳng phải khó. Bạn sẽ ngạc nhiên khi thấy bao nhiêu kết luận không có cơ sở được rút ra ngay sau khi vấn đề không phải là ngớ ngẩn. Con người dường như có một sự ưu ái đặc biệt dành cho những giải thích nhân-quả đơn giản.

Trẻ em có bàn chân lớn hơn thì viết tốt hơn so với trẻ em có bàn chân nhỏ: kích cỡ bàn chân là yếu tố quan trọng đóng góp vào kỹ năng viết? Không phải đâu, đó là vì trẻ em có bàn chân lớn hơn thường là lớn tuổi hơn và đã được luyện viết nhiều hơn.

Nhiều mối tương quan có thể hoàn toàn ngẫu nhiên. Cùng lúc với nhiệt độ toàn cầu tăng lên, số cướp biển cũng vậy. Cái đó thì dễ nhận ra là tình cờ. Nhưng nếu đó không phải là số cướp biển mà là số cơn bão, số lượng xe hơi, số đồ bơi được bán, hay quần thể thằn lằn? Có lẽ không còn ngẫu nhiên nữa, bởi chúng có mối quan hệ với nhiệt độ toàn cầu.

Một trường hợp đặc biệt là các yếu tố nguy cơ trong y học. Huyết áp cao và cholesterol cao là những yếu tố nguy cơ gây bệnh tim, vì vậy mọi người được khuyên nên ăn uống lành mạnh. Nhưng lượng cholesterol cao không phải là nguyên nhân gây ra bệnh tim! Đó chỉ là vì người ta thấy tỉ lệ mắc bệnh tim trong nhóm người có hàm lượng cholesterol cao hơn so với nhóm có hàm lượng cholesterol bình thường. Và vì ngày càng có nhiều người có lượng cholesterol bình thường hơn, chúng ta thấy hầu hết những người chết vì bệnh tim đều có mức cholesterol hoàn toàn bình thường.

Đi bộ với một cây gậy là một yếu tố nguy cơ gây té ngã. Dữ liệu thống kê cho thấy có nhiều người chống gậy bị ngã hơn những người không chống gậy! Nhưng thử bỏ cây gậy – yếu tố nguy cơ – sang một bên xem: mọi thứ chỉ có thể tồi tệ hơn.



Ngày nay, “ý nghĩa thống kê” (statistical significance) như là thuật ngữ thần diệu trong khoa học. Cách đây không lâu, không ai nghĩ về ý nghĩa; nhưng bây giờ mọi thứ phải “có ý nghĩa” để được để ý, thậm chí được xuất bản trong các tạp chí khoa học. Mặc dù vậy, bản thân hầu hết các nhà khoa học chưa chắc đã hiểu một kết quả có ý nghĩa thống kê” thực sự có nghĩa là gì. Nó là một khái niệm khó.



### 9.9 Ý nghĩa thống kê

“Ý nghĩa” là cách khoa học tách các kết quả thật (xảy ra một cách có hệ thống) từ những phát hiện hoàn toàn do tình cờ. Nếu một nhà khoa học có một phát hiện có ý nghĩa, những gì ông/bà ấy thực sự muốn nói là: "Kết quả này là thật vì nó rất khó xảy ra do ngẫu nhiên được."

**Kiểm định giả thuyết** là kiểm định ý nghĩa thống kê thông qua việc đặt ra hai giả thuyết phủ định nhau và làm một thí nghiệm để xem cái nào là phù hợp với thực tế (tức là nó không xảy ra một cách tình cờ).

Chẳng hạn, tôi đưa giả thuyết: học viên khóa này thông minh hơn so với người dân nói chung (IQ bình quân = 100, độ lệch chuẩn = 15). Để kiểm định giả thuyết này, tôi đưa ra một cái gọi là "giả thuyết vô hiệu" (null hypothesis) để phủ định giả thuyết ban đầu: chỉ số IQ trong nhóm học viên và trong người dân nhà là như nhau (không có sự khác biệt nào cả).

Nếu kết quả thử nghiệm cho thấy rõ ràng là xác suất để IQ hai nhóm bằng nhau là rất thấp, thì sự bằng nhau đó “không có ý nghĩa thống kê”. Tức là trong thực tế, nếu tìm thấy hai IQ hai nhóm này bằng nhau, thì đó là do ngẫu nhiên mà thôi. Như thế, tôi có thể kết luận rằng giả thuyết ban đầu không thể xảy ra do tình cờ mà là rất phù hợp với thực tế. Nói cách khác, kết quả kiểm định giả thuyết ban đầu “có ý nghĩa thống kê”.

Để kiểm định, vì chúng ta không yêu cầu tất cả học viên làm bài kiểm tra chỉ số IQ mà chỉ chọn ngẫu nhiên một số người trong số đó để tạo nên một mẫu đại diện cho tất cả họ. Chúng ta có thể chọn rất nhiều mẫu ngẫu nhiên theo nhiều cách khác nhau, nhưng nếu đại đa số các mẫu học viên này đều cho kết quả IQ bình quân khác với IQ bình quân trong cộng đồng, thì chúng ta sẽ kết luận là giả thuyết ban đầu có ý nghĩa thống kê và phù hợp với thực tế.

Hãy giả sử rằng tôi đã chọn ra một mẫu học viên đại diện để làm bài kiểm tra IQ, và họ đạt điểm IQ bình quân là 105, với độ lệch chuẩn là 2 điểm. Rõ ràng, một số đạt điểm cao hơn một chút so với 105, một vài người thì thấp hơn một chút. Nghĩa là, có một sự phân tán trong tập hợp các điểm số IQ.

Một lần nữa, hãy nhớ là 95% tập hợp dữ liệu sẽ nằm trong 4 độ lệch chuẩn. Trong trường hợp này, nếu chọn hết mẫu này đến mẫu khác từ tất cả học viên và tập hợp các giá trị IQ bình quân từ các mẫu này lại, chúng ta sẽ thấy 95% nằm trong 101 đến 109. Đây gọi là khoảng tin cậy (confidence interval): chúng ta tin rằng 19 trong số 20 mẫu tương tự sẽ có chỉ số IQ trong khoảng giữa 101 và 109.

Chỉ có 5% mẫu học viên (1 trong 20) có chỉ số IQ bình quân nằm ngoài khoảng này. Nói cách khác, xác suất cao nhất để chỉ số IQ học viên bằng 100 (tức bằng IQ cộng đồng) là 5%. Đó là một tỷ lệ rất nhỏ -- thậm chí quá nhỏ -- và chúng ta có thể xem một kết quả như vậy là do cơ hội tình cờ. Vì vậy, chúng ta có thể bác bỏ giả thuyết vô hiệu ở trên, để từ đó xác nhận rằng giả thuyết ban đầu là phù hợp thực tế. Tóm lại, đúng là học viên khóa học này có chỉ số IQ cao hơn so với phần cộng đồng dân chúng nói chung – một cách có ý nghĩa thống kê.

Một lỗi thường thấy trong giới nhà báo khoa học – và rất nhiều nhà khoa học – là **nhầm lẫn ý nghĩa thống kê với tầm quan trọng thực tế**. Một cái gì đó có ý nghĩa thống kê chưa hẳn là có tầm quan trọng thực tế. Nó chỉ đơn giản có nghĩa là xác suất để điều đó không xảy ra tình cờ là rất cao.

Giả sử một nhà khoa học thử nghiệm công dụng hai loại thuốc tăng trọng mới trên trẻ em. Họ tìm ra rằng khả năng tăng trọng ở nhóm uống thuốc A so với nhóm đối chứng là 33% và nhóm uống thuốc B so với nhóm đối chứng là 35%. Kiểm định giả thuyết cho thấy sự khác nhau này là có ý nghĩa thống kê (trong 95% trường hợp, nhóm trẻ thuốc B sẽ tăng trọng nhiều hơn nhóm trẻ thuốc A).

Nhưng nếu nhìn vào hai con số, sự khác biệt 2 điểm phần trăm này chưa chắc là đã thực sự quan trọng. Nói cách khác, hiệu quả tăng trọng từ thuốc A và thuốc B có khác biệt theo ý nghĩa thống kê, nhưng trong thực tế, nhiều phụ huynh có thể sẽ cho rằng chúng gần như nhau về hiệu quả và mua thuốc nào cũng được.



Cách dễ dàng nhất để tạo ra những số liệu thống kê “lừa bịp”, buồn thay, cũng đang phổ biến rộng rãi nhất: bỏ qua một nửa câu chuyện. Nói cách khác, ở đây, chúng ta đã đi đủ vòng tròn: tất cả đều là vấn đề về bối cảnh. Các nhà báo không chỉ phải cẩn trọng với những thông tin mà họ được cung cấp, mà cả những thông tin họ không được cung cấp.



### 9.10 Vài cạm bẫy cuối cùng

Quay trở lại với ví dụ về loại thuốc chống trầm cảm – 40% người sử dụng thuốc sẽ gặp sự cố tình dục. Câu hỏi thông minh nhất đã bị cho ra rìa: có bao nhiêu người không sử dụng vẫn gặp vấn đề khi giao hợp? Nếu 30% trong số đó cũng gặp sự cố tình dục thì mọi chuyện có vẻ hoàn toàn khác, vì nó cho thấy sự khác biệt chỉ ở mức 10 điểm phần trăm.

Việc ghi nhớ nhóm đối chứng (control group) để so sánh với nhóm thử nghiệm (hay nhóm điều trị trong y khoa) luôn luôn rất quan trọng – không có nhóm đối chứng thì không phải là khoa học.

Chúng ta đọc trên báo rằng một loại gien làm tăng gấp đôi nguy cơ bị ADHD (bệnh thiếu chú ý và quá hiếu động ở trẻ em). Nhưng vấn đề là có bao nhiêu trẻ em bị ADHD có gien này, và bao nhiêu không có nó? Sẽ là tỉ lệ 2 trên 100 so với 1 trên 100? Điều đó sẽ không đủ để hậu thuẫn cho tuyên bố rằng “các gien gây ADHD đã được tìm thấy”. Nếu đó là 7% so với 14% thì sao?

Đây là một cái bẫy khác. Nếu bất lực là rất hiếm, chỉ ở mức 1 trong 100 (lần giao hợp), một sự gia tăng 10% tỷ lệ đó (trở thành 1 trong 91 lần giao hợp) là có thể chịu đựng được. Chỉ khi tỉ lệ ban đầu là cao (xảy ra rất thường xuyên) thì mức tăng 10% mới có thể phá hỏng các mối quan hệ vợ chồng. Tỉ lệ ban đầu là rất quan trọng khi đánh giá nguy cơ thực tế. Trong các vấn đề khác cũng vậy: những cái rất hiếm thì rất khó tìm.

Ngay cả khi chúng không quá hiếm, một thí nghiệm có thể cho ra nhiều kết quả “dương tính giả” (thử nghiệm thấy có bệnh nhưng thực ra là khỏe mạnh) hơn “dương tính thật”, đơn giản chỉ vì có quá nhiều người khỏe mạnh. Tương tự như vậy, thử nghiệm hàng loạt để tìm khủng bố ở sân bay là không khả thi: bạn sẽ lọc ra nhiều người dân vô tội hơn những kẻ khủng bố. Vấn đề tương tự cũng sẽ xảy ra khi sàng lọc hàng loạt để tìm ung thư vú, ung thư tuyến tiền liệt... Tóm lại, **tìm một cây kim trong đồng cỏ khô là điều không thể, không phải vì bạn sẽ không tìm thấy cây kim, mà vì trong đầu bạn cứ nghĩ đã tìm thấy nó.**

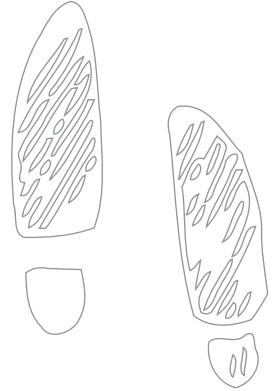


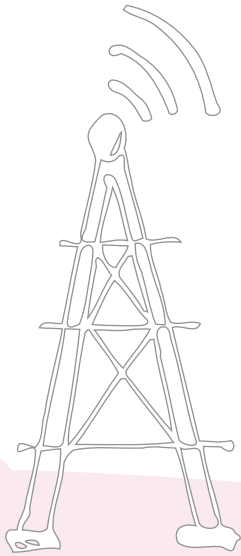


### 9.11 Kết luận

Chúng ta sẽ kết thúc với một lời cảnh báo. Có lẽ, không lĩnh vực khoa học nào lại đầy rẫy những khó khăn, giả định và bẫy ẩn như thống kê. Có vẻ như nó quá dễ để mắc một sai lầm.

Vì vậy, ngay cả khi bạn đã xong khóa học này, đừng cố gắng trở thành một thống kê nghiệp dư. **Khi có bất cứ nghi ngờ nào, hãy gọi một nhà thống kê.** Và nếu bạn chưa biết một người nào, hãy tìm và kết bạn với một ai đó mà bạn có thể gọi bất cứ lúc nào, bất kể ngày hay đêm. Tôi thấy từ kinh nghiệm rằng giới thống kê là nhóm nhà khoa học hữu ích nhất.





### 9.12 Câu hỏi tự luận (1-6)

1. Ngôi sao gần nhất chúng ta là Proxima Centauri, cách 4,2 năm ánh sáng. Một năm ánh sáng là 10 nghìn tỷ km. Cố gắng giảm thiểu con số đó đến phạm vi con người – nhưng đừng nói “ánh sáng từ ngôi sao phải mất 4,2 năm để đến được chúng ta”!
2. Theo các hãng thông tấn, 200.000 người đến để nghe Barack Obama ở Berlin vào năm 2008. Liệu chúng ta có thể đưa tin như vậy?
3. Tổng diện tích lá phổi theo một đơn vị để hiểu là bao nhiêu?
4. Sinh viên phải trả thêm một phần ba cho các khoản vay học phí: lãi suất tăng từ 2,7% lên 3,7%. Bạn sẽ viết tin này như thế nào?
5. 1/3 của 27% của 405 là bao nhiêu?
6. Cà phê đã khử chất cafein khiến lượng Apo B (protein kết hợp lipid) trong máu tăng 8% (<http://www.sciencedaily.com/releases/2005/11/051120122949.htm>). Những con số nào bạn cần biết trước khi quyết định làm tin này? Còn gì để nói về câu chuyện này?



### 9.13 Đáp án câu hỏi tự luận (1-6)

1. Ngôi sao gần nhất chúng ta là Proxima Centauri, cách 4,2 năm ánh sáng. Một năm ánh sáng là 10 nghìn tỷ km. Cố gắng giảm thiểu con số đó đến phạm vi con người – nhưng đừng nói “ánh sáng từ ngôi sao phải mất 4,2 năm để đến được chúng ta”!

**ĐÁP ÁN:** Nếu trái đất được đại diện bởi một dấu chấm ở cuối câu này (0,5 mm), ngôi sao gần nhất sẽ cách xa 1570 km – gần bằng khoảng cách đường bộ từ Sài Gòn đi Hà Nội (khoảng 1700 km).

2. Theo các hãng thông tấn, 200.000 người đến để nghe Barack Obama ở Berlin vào năm 2008. Liệu chúng ta có thể đưa tin như vậy?

**ĐÁP ÁN:** Trong một đám đông, chúng ta ước tính có 2-4 người trên một mét vuông. Do đó, để chứa đám đông 200.000 người, sẽ cần một diện tích kích cỡ chùng 300 x 300 mét (tức chùng 12 sân bóng đá đặt lại với nhau). Tuy nhiên, đại lộ nơi Obama nói chuyện ở Berlin chỉ rộng 50 mét...

3. Tổng diện tích lá phổi theo một đơn vị dễ hiểu là bao nhiêu?

**ĐÁP ÁN:** Tổng diện tích bề mặt phổi ở người trưởng thành là 70 mét vuông. Hãy nói nó bằng một nửa sân tennis.

4. Sinh viên phải trả thêm một phần ba cho các khoản vay học phí: lãi suất tăng từ 2,7% lên 3,7%. Bạn sẽ viết tin này như thế nào?

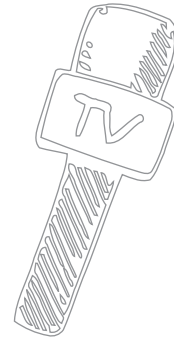
**ĐÁP ÁN:** Đưa tin về sự khác biệt theo đồng tiền thực tế -- hoặc tốt hơn là ý nghĩa của nó đối với thu nhập sinh viên.

5. 1/3 của 27% của 405 là bao nhiêu?

**ĐÁP ÁN:**  $1/3 \times 27/100 \times 405 = (27 \times 405) / (3 \times 100) = 36.5$ . Đừng đưa quá nhiều số thập phân vào trong kết quả: Bản thân 1/3 cũng chẳng phải là con số chính xác!

6. Cà phê đã khử chất cafein khiến lượng Apo B (protein kết hợp lipid) trong máu tăng 8%. Những con số nào bạn cần biết trước khi quyết định làm tin này? Còn gì để nói về câu chuyện này?

**ĐÁP ÁN:** Tìm ra giá trị "bình thường" của Apo B trong cộng đồng nói chung; tìm hiểu xem nếu tăng một lượng Apo B có làm tăng nguy cơ mắc bệnh; tìm hiểu xem liệu con số đó chỉ là một một trong nhiều phát hiện có thể tương tự khác từ nghiên cứu này...





### 9.14 Câu hỏi tự luận (7-12)

7. Vào năm 1982, nhà khảo cổ học nổi tiếng Stephen Jay Gould được chẩn đoán có một khối u, và các bác sĩ nói với ông thời gian sống sót trung vị là tám tháng. Vì vậy, SJG khá lạc quan về cơ hội của mình. Tại sao?
8. Thời gian chờ đợi bình quân tại hai phòng khám bác sĩ là 30 phút. Một cái có độ lệch chuẩn là 0 phút, cái còn lại là 15 phút. Vậy bạn muốn bác sĩ nào hơn?
9. Việc sàng lọc ung thư vú bằng cách chụp nhũ ảnh làm giảm nguy cơ tử vong cho phụ nữ tầm sáu mươi tuổi đến 25%. Có khoảng 4 trong 1000 phụ nữ trong nhóm tuổi này chết vì ung thư vú. Bạn có thể diễn đạt lại tỷ lệ phần trăm trên theo cách khác? Con số nào cung cấp một ý tưởng tốt hơn về những gì đang xảy ra?
10. Số lượng nữ bộ trưởng trong nội các chính phủ mới đã tăng gấp ba. Đó có phải là một tin tốt?
11. Chúng ta biết mối tương quan không có nghĩa là quan hệ nhân quả -- hai hiện tượng có liên hệ, không có nghĩa là cái này gây ra cái kia. Thực hành nguyên tắc này với các câu hỏi sau:
- a. Trẻ thụ tinh ống nghiệm có nhiều biến chứng lúc sinh hơn những trẻ thụ tinh tự nhiên. Phải chăng là việc thụ tinh trong ống nghiệm gây nguy hiểm cho các ca sinh hơn?
  - b. Người cao hơn kiếm được nhiều hơn so với những người thấp hơn. Chiều cao làm nên thu nhập?
  - c. Trong 95% các cuộc gây hấn chết người tại các quán bar, nạn nhân là người gây nên mọi sự. Bạn nghĩ gì?
12. Liệu người thợ kim hoàn có thực sự sử dụng vàng 18 carat trên chiếc vương miện ông làm cho nhà vua? Archimedes được lệnh phải tìm ra điều này. Ông tìm được rằng mật độ vàng 18 carat là 15,5 gram trên một cen-ti-mét khối, Trên chiếc vương miện, Archimedes tìm thấy mật độ vàng là 15,1 gram mỗi cm khối. Khá gần – nhưng ông ta vẫn khuyên nhà vua sa thải người thợ kim hoàn. Lý do tại sao?





## 9.15 Đáp án câu hỏi tự luận (7-12)

7. Vào năm 1982, nhà khảo cổ học nổi tiếng Stephen Jay Gould được chẩn đoán có một khối u, và các bác sĩ nói với ông thời gian sống sót trung vị là tám tháng. Vì vậy, SJG khá lạc quan về cơ hội của mình. Tại sao?

**ĐÁP ÁN:** Một nửa số người mắc bệnh chết trong vòng tám tháng, nhưng có thể trong một nửa trên tám tháng còn lại, vẫn có những người giữ được tuổi thọ bình thường của mình, đặc biệt là nếu ung thư được chẩn đoán sớm, đi kèm với các yếu tố thuận lợi khác. Gould là trường hợp hy hữu trong nhóm “trên tám tháng”, sống đến 20 năm nữa nhờ điều trị, thái độ tích cực với bệnh, và nhiều yếu tố khác.

8. Thời gian chờ đợi bình quân tại hai phòng khám bác sĩ là 30 phút. Một cái có độ lệch chuẩn là 0 phút, cái còn lại là 15 phút. Vậy bạn muốn bác sĩ nào hơn?

**ĐÁP ÁN:** Nếu độ lệch chuẩn là 0, bạn luôn luôn chờ đợi đúng 30 phút. Nếu độ lệch chuẩn là 15 phút, đôi khi bạn có thể đi thẳng vào gặp, đôi khi bạn sẽ phải chờ đợi một giờ đồng hồ. (Hãy nhớ 95% trường hợp sẽ nằm trong khoảng 4 độ lệch chuẩn – trong trường hợp này là  $30-2 \times 15=0$  phút đến  $30+2 \times 15=60$  phút.)

9. Việc sàng lọc ung thư vú bằng cách chụp nhũ ảnh làm giảm nguy cơ tử vong cho phụ nữ tầm sáu mươi tuổi đến 25%. Có khoảng 4 trong 1000 phụ nữ trong nhóm tuổi này chết vì ung thư vú. Bạn có thể diễn đạt lại tỷ lệ phần trăm trên theo cách khác? Con số nào cung cấp một ý tưởng tốt hơn về những gì đang xảy ra?

**ĐÁP ÁN:** Nếu không có việc sàng lọc, 4 trong 1000 phụ nữ sẽ chết sớm vì bệnh ung thư vú. Với việc sàng lọc, tỉ lệ này còn 3 trong 1000. Rõ ràng, nói như thế sẽ dễ cho người đọc hình dung hơn vì mức giảm nguy cơ tử vong nhờ sàng lọc.

10. Số lượng nữ bộ trưởng trong nội các chính phủ mới đã tăng gấp ba. Đó có phải là một tin tốt?

**ĐÁP ÁN:** Sự thay đổi là chậm nếu nó tăng từ 1 lên 3, nhưng sẽ là đáng kể nếu nó tăng từ 6 lên 18 – đặc biệt là nếu có 18 bộ trưởng trong nội các!

11a. Trẻ thụ tinh ống nghiệm có nhiều biến chứng lúc sinh hơn những trẻ thụ tinh tự nhiên. Phải chăng là việc thụ tinh trong ống nghiệm gây nguy hiểm cho các ca sinh hơn?

**ĐÁP ÁN:** Nói chung, các bà mẹ thụ tinh ống nghiệm thường lớn tuổi hơn và việc thụ tinh ống nghiệm dẫn đến nhiều ca đa sinh (sinh đôi, sinh ba...) hơn. Cả hai yếu tố này (lớn tuổi và đa sinh) đều có mối tương quan với các biến chứng sinh sản. Nếu hai yếu tố này được điều chỉnh trong mô hình thống kê thích hợp, sẽ không có khác biệt gì về nguy cơ biến chứng lúc sinh giữa trẻ thụ tinh trong ống nghiệm và thụ tinh thường. Trên thực tế, giới khoa học phải mất một thời gian khá khá mới giải thích được vấn đề này.

11b. Người cao hơn kiếm được nhiều hơn so với những người thấp hơn. Chiều cao làm nên thu nhập?

**ĐÁP ÁN:** Đàn ông, về trung bình, cao hơn so với phụ nữ, và kiếm được nhiều tiền hơn phụ nữ. Đó có thể là một lý do cho mối tương quan giữa chiều cao và thu nhập.

Tuy nhiên, nếu chúng ta bỏ qua vấn đề giới tính, chúng ta vẫn tìm thấy một sự khác biệt: đàn ông cao hơn kiếm được nhiều tiền hơn so với đàn ông thấp hơn, và phụ nữ cao hơn kiếm được nhiều hơn so với phụ nữ thấp hơn. Bạn biết lý do tại sao không?

11c. Trong 95% các cuộc gây hấn chết người tại các quán bar, nạn nhân là người gây nên mọi sự. Bạn nghĩ gì?

**ĐÁP ÁN:** Trong những trận gây hấn chết người – các báo cáo cảnh sát cho thấy – một trong những người bắt đầu gây chiến thường kết thúc bằng cái chết. Nhưng đó chỉ là vì người sống sót có cơ hội để nói ai là người bắt đầu gây sự!

12. Liệu người thợ kim hoàn có thực sự sử dụng vàng 18 carat trên chiếc vương miện ông làm cho nhà vua? Archimedes được lệnh phải tìm ra điều này. Ông tìm được rằng mật độ vàng 18 carat là 15,5 gram trên một cen-ti-mét khối, Trên chiếc vương miện, Archimedes tìm thấy mật độ vàng là 15,1 gram mỗi cm khối. Khá gần – nhưng ông ta vẫn khuyên nhà vua sa thải người thợ kim hoàn. Lý do tại sao?

**ĐÁP ÁN:** Bởi vì Archimedes thấy mật độ vàng trên mỗi cm khối ở vương miện khác 15,5 gram một cách “có ý nghĩa thống kê”! Ông có thể đã đo mật độ vàng trên vương miện nhiều lần để chắc chắn, và mỗi lần cho ra một mật độ khác nhau. Tập hợp các mật độ đó có một độ lệch chuẩn là 0,15 gram mỗi cm khối. Nếu mật độ bình quân là 15,1 gram trên cm khối thì khoảng tin cậy ở mức 95% là từ  $15,1-2 \times 0,15=14,8$  đến  $15,1+2 \times 0,15=15,4$  gram mỗi cm khối. Khoảng tin cậy này vẫn nằm dưới mật độ 15,5 gram mà người thợ kim hoàn lẽ ra phải làm cho chiếc vương miện.



## 9.16 Bài thực hành

1. Hãy tìm trên Internet hay liên lạc với Tổng cục Thống kê xem bạn có thể có được dữ liệu về phân bổ thu nhập ở Việt Nam. Nếu có, hãy phân tích xem thu nhập phân bổ thế nào, bao nhiêu người trên mức bình quân, mức trung vị là bao nhiêu... Kết luận là gì?

2. Hãy vào số liệu về chi ngân sách nhà nước VN giai đoạn 2005-2010 [<http://www.gso.gov.vn/default.aspx?tabid=389&idmid=3&ItemID=12947>] và phân tích xem mức chi ngân sách cho hai lĩnh vực y tế và lĩnh vực khoa học-công nghệ-môi trường tăng hay giảm thế nào qua mỗi năm và trong toàn thời kỳ năm năm này. Hãy thử viết một bản tin ngắn 300-400 từ dựa trên kết quả vừa phân tích và thảo luận nó với người hướng dẫn. Hãy ghi nhớ hai khái niệm phần trăm và điểm phần trăm.

3. Hãy vào trang web thống kê của Tổ chức Y tế Thế giới (WHO): [<http://www.who.int/topics/statistics/en/>]. Tìm các số liệu y tế mới nhất về Việt Nam và thảo luận với cố vấn xem những số liệu thống kê nào trong số đó có thể là nguồn cảm hứng cho một bài báo.

4. Hãy xem buổi diễn thuyết có tựa "Những con số liệu thống kê sẽ định hình lại nhân quan thế giới của bạn" tại: [[http://www.ted.com/talks/hans\\_rosling\\_shows\\_the\\_best\\_stats\\_you\\_ve\\_ever\\_seen.html](http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html)] Nhà thống kê Hans Rosling bóc trần những huyền thoại về cái gọi là "thế giới đang phát triển". Sau đó, vào website mà ông đề cập đến, [<http://www.gapminder.org>].

Tìm chung quanh website này càng nhiều số liệu thống kê càng tốt về Việt Nam và các nước láng giềng. Thảo luận những phát hiện của bạn với người cố vấn. Sau đó, hãy thử làm một bài ký hay bản tin nào đó từ những số liệu thu thập được.

5. Hãy trang web sau, nơi tập hợp một số bài báo tại Việt Nam về tác dụng tích cực và tiêu cực của Aspirin: [<http://www.khoahoc.com.vn/timkiem/Aspirin/index.aspx>]. Nếu đường link không hoạt động, hãy vào trang chủ webstie này và đánh từ khóa aspirin vào ô tìm kiếm.

Hãy phân tích và viết một bài bình luận ngắn về cách đưa tin bài về tác dụng dược phẩm trên báo chí, lấy aspirin làm một điển hình. Chú ý đặc biệt đến các con số về mức độ tăng giảm nguy cơ trong các bài báo. Nếu cần, hãy tìm thêm số liệu hoặc tư liệu gốc, phỏng vấn thêm chuyên gia trong nước... để phân tích, kiểm chứng số liệu hay thu thập ý kiến. Trao đổi về tác phẩm cuối cùng với người hướng dẫn.

6. Ghé thăm sáu trang blog thống kê dưới đây một cách thường xuyên. Bạn có thể tìm được nhiều cảm hứng để vạch trần một số huyền thoại thống kê nào đó!

Freakonomics	<a href="http://freakonomics.blogs.nytimes.com/">http://freakonomics.blogs.nytimes.com/</a>
Bad Science	<a href="http://www.badsience.net/">http://www.badsience.net/</a>
Junk Charts	<a href="http://junkcharts.typepad.com/junk_charts/">http://junkcharts.typepad.com/junk_charts/</a>
Stats Blog	<a href="http://thestatsblog.wordpress.com/">http://thestatsblog.wordpress.com/</a>
The Numbers Guy	<a href="http://blogs.wsj.com/numbersguy/">http://blogs.wsj.com/numbersguy/</a>
Statistical Modeling, Causal Inference, and Social Science	<a href="http://www.stat.columbia.edu/~gelman/blog/">http://www.stat.columbia.edu/~gelman/blog/</a>