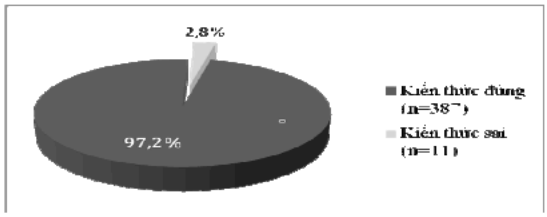
Những nguyên tắc trình bày biểu đồ trong bài báo khoa học

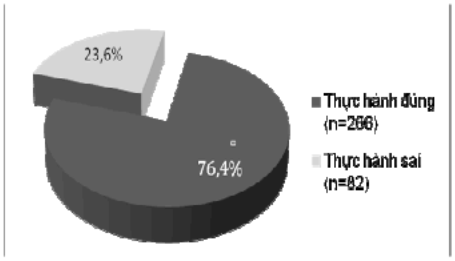
Nguyễn Văn Tuấn

Một trong những vấn đề hay thấy trong các bài báo khoa học ở Việt Nam là cách trình bày dữ liệu bằng biểu đồ. Những biểu đồ được thiết kế quá đơn giản (phần lớn là cắt và dán từ các phần mềm máy tính) và vi phạm hầu như bất cứ nguyên tắc nào của trình bày dữ liệu mà có lẽ tác giả chưa làm quen. Trong chương này, tôi sẽ bàn qua những nguyên tắc trình bày dữ liệu trong biểu đồ.

Để cảm nhận được vấn đề, có thể xem qua vài biểu đồ hay thấy trong các bài báo khoa học ở Việt Nam:

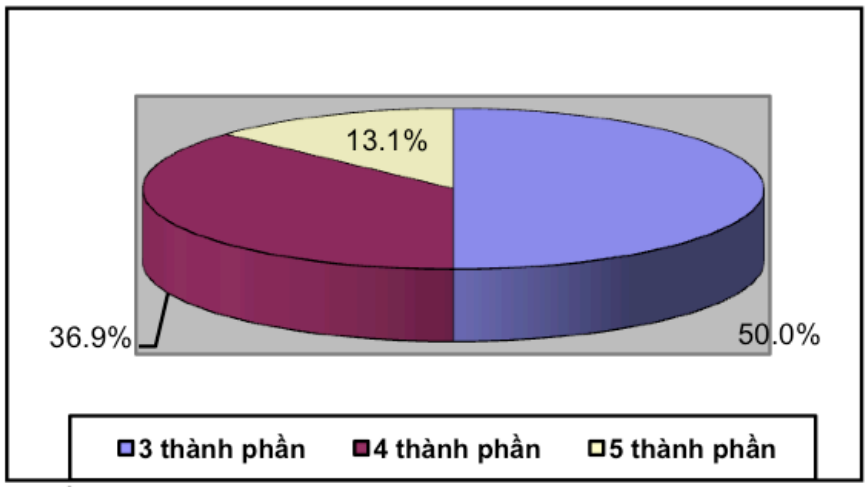


Hình 1: *Kiến thức chung về phòng tránh té ngã*



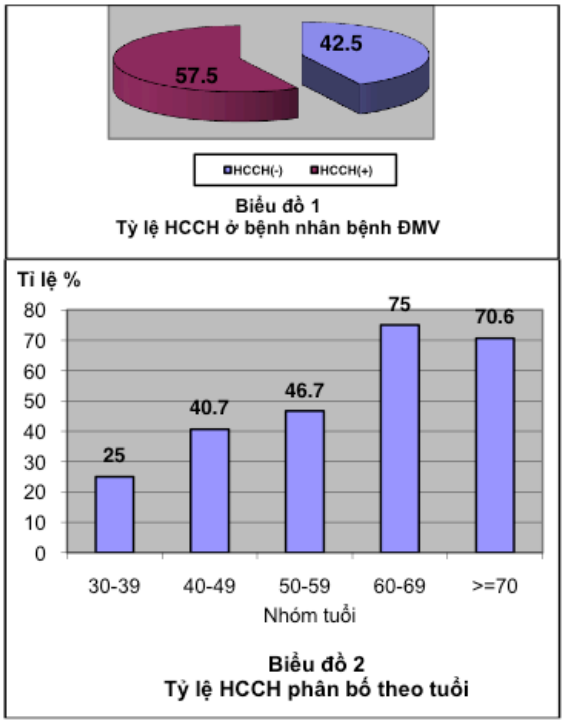
Hình 3: *Thực hành chung về phòng tránh té ngã*

Trên đây là biểu đồ mô tả kết quả. Mỗi biểu đồ thật ra chỉ có 2 con số!

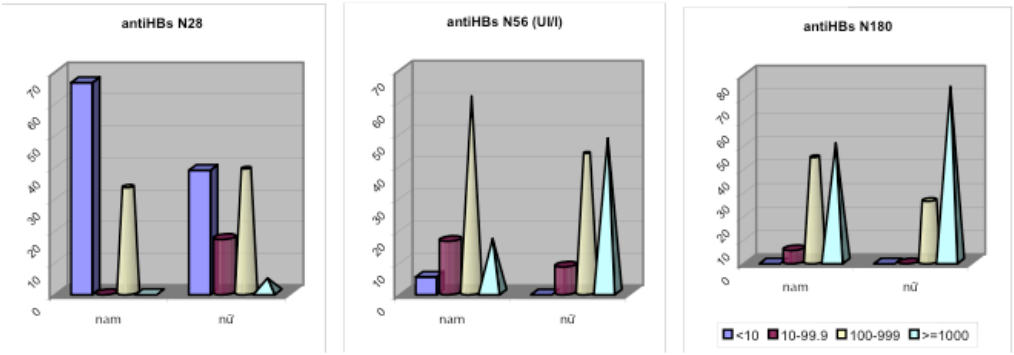


***Biểu đồ 7**: Tỷ lệ số thành phần trong HCCH*

Biểu đồ này chẳng những có thể xem là nghèo nàn vì ít số liệu, mà còn tốn khá nhiều mực in và cả màu.



Hai biểu đồ trên cũng nghèo nàn về số liệu, và lượng thông quá thấp.



Biểu đồ này rất khó đọc, một phần là do phông chữ quá nhỏ, một phần là do chọn các bar để thể hiện dữ liệu.

Những biểu đồ như trình bày trên thật ra khá phổ biến trên các tập san khoa học trong nước. Có thể nói rằng phần lớn biểu đồ không cung cấp thông tin liên quan hay thông tin quan trọng để bổ sung thông điệp chính của bài báo. Đại đa số biểu đồ được soạn một cách hời hợt, làm cho ngươi đọc cảm thấy tác giả hình như chưa đầu tư vào việc suy nghĩ và thiết kế. Thật ra, trong thực tế, phần lớn các biểu đồ trên các tập san khoa học ở VN là cắt và dán trực tiếp từ các phần mềm như Excel hay phần mềm thống kê. Chính vì thế mà khi đọc những biểu đồ, có nhiều kí hiệu, cách diễn tả rất khó hiểu (lẫn lộn giữa tiếng Anh và tiếng Việt). Một bài báo khoa học với những biểu đồ như thế rất khó có cơ may được chấp nhận cho công bố trên các tập san quốc tế.

Nguyên tắc soạn biểu đồ

Biểu đồ là một cách trình bày dữ liệu khoa học rất hữu hiệu. Người Trung Hoa từng có câu *một hình ảnh có giá trị tương đương với một vạn chữ.* Thật vậy, đối phó với một rừng số liệu thu thập từ thí nghiệm, vấn đề đặt ra là làm sao khai thác những số liệu này một cách hữu hiệu nhất. Hữu hiệu ở đây phải hiểu là chuyển tải thông tin sao cho cho người xem cảm thấy dễ lĩnh hội nhất. Có ba cách

để thể hiện dữ liệu khoa học: dùng chữ viết, bảng số liệu, và biểu đồ. Chữ viết chỉ có thể sử dụng cho những dữ liệu rất đơn giản, chứ không thể hiện được tất cả những xu hướng và dao động của dữ liệu. Bảng số liệu có thể sử dụng cho trường hợp tóm lược những thông tin mang tính chính xác cao. Nhưng biểu đồ có thể sử dụng để chuyển tải những thông điệp về mức độ ảnh hưởng và xu hướng biến thiên của dữ liệu. Do đó, đứng trước quyết định chọn hình thức để thể hiện dữ liệu, biểu đồ phải và nên xem là một hình thức số 1.

Để thiết kế biểu đồ một cách hữu hiệu, cần phải làm quen với cái tên "Edward Tufte", vì ông là một chuyên gia hàng đầu về biểu đồ. Ông là giáo sư thống kê học của Đại học Yale, giáo sư chính trị học, và giáo sư khoa học máy tính (cũng tại Yale), là người đã có ảnh hưởng cực kì lớn đến lĩnh vực trình bày dữ liệu bằng biểu đồ, qua những công trình có thể nói là đặt nền tảng cho lĩnh vực này (có khi được đề cập đến như là *data visualization*). Ông là người dám thuế chấp căn nhà mình cho ngân hàng để vay một số tiền làm nghiên cứu và cho ra công trình về *data visualization* mà sau này ông không bao giờ hối hận (vì quá thành công về tài chính!) Báo *New York Times* gọi ông là *Leonardo Da Vinci of Data.*

Edward Tufte đặt ra triết lí và 4 nguyên tắc trong trình bày dữ liệu bằng biểu đồ. Triết lí của thể hiện dữ liệu có thể tóm lược trong câu sau đây: "Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space" (tạm dịch: *triết lí của trình bày dữ liệu bằng biểu đồ là cung cấp cho người xem một lượng thông tin cao nhất trong một thời lượng nhỏ nhất với lượng mực in thấp nhất trong một không gian nhỏ nhất*). Như vậy, khi trình bày dữ liệu bằng biểu đồ, cần phải chú ý đến 4 khía cạnh: lượng thông tin, thời gian, lượng mực in, và không gian. Triết lí này có thể xem như là một triết lí hà tiện, tốn ít tài lực nhất để có nhiều thông tin nhất. Để đạt được triết lí đó, Tufte đặt ra 4 nguyên tắc:

- Nói lên sự thật về dữ liệu;
- Tối đa hoá tỉ số dữ liệu trên mực in;
- Tối đa hoá mật độ dữ liệu; và
- Trình bày dữ liệu một cách đầy đủ, không phải trang trí biểu đồ.

Một số biểu đồ minh họa dưới đây được trích từ sách của ông Tufte.

Yếu tố dối (Lie factor)

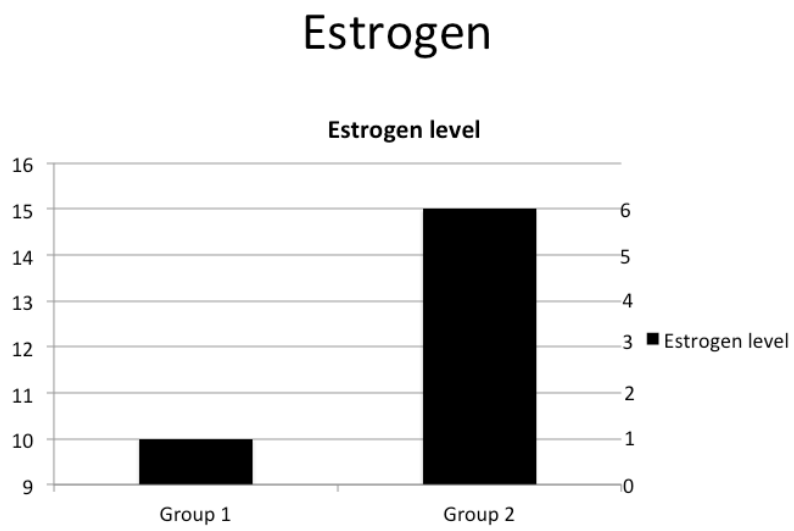
Tufte khuyến cáo rằng việc thể hiện số liệu trên nền của biểu đồ phải theo tỉ lệ thuận với định lượng của trục tung và trục hoành. Nếu số liệu cho thấy tỉ lệ tăng trưởng là 30%, và biểu đồ cũng thể hiện con số đó, thì không có vấn đề gì. Nhưng

nếu biểu đồ được thiết kế làm cho mức độ ảnh hưởng lớn hơn mức độ thật thì đó là một sự gian dối. Do đó, Tufte định nghĩa Lie factor (tạm dịch: Yếu tố dối) là tỉ số của mức độ ảnh hưởng trình bày trên biểu đồ với mức độ ảnh hưởng của số liệu. Gọi LF là yếu tố dối, định nghĩa này có nghĩa là:

$$LF = ESgraph / ESdata$$

Trong đó, ESgrap là mức độ ảnh hưởng của biểu đồ (effect size in graph), và ESdata là mức độ ảnh hưởng của số liệu (effect size in data). Biểu đồ nên được thiết kế sao cho LF gần bằng 1. Nói cách khác, LF càng cao thì mức độ nói dối càng cao.

Chúng ta có thể lấy một ví dụ sau đây để làm ví dụ. Trong biểu đồ dưới đây, tác giả trình bày nồng độ estrogen cho hai nhóm (Group 1 và Group 2). Nhìn qua biểu đồ, chúng ta có lẽ rất ấn tượng vì nồng độ estrogen có vẻ rất khác biệt giữa hai nhóm bệnh nhân. Nhưng nếu nhìn kĩ, chúng ta thấy có rất nhiều vấn đề trong biểu đồ này, nhưng chúng ta bàn qua yếu tố dối trước.



Để xem yếu tố dối, chúng ta cần phải tính mức độ ảnh hưởng của dữ liệu. Chú ý rằng Nhóm 1 có nồng độ estrogen là 10, và nhóm 2 là 15. Do đó, mức độ ảnh hưởng có thể tính bằng cách lấy giá trị cao nhất trừ cho giá trị thấp nhất, và chia kết quả cho giá trị thấp nhất:

$$ESdata = (15 – 10) / 10 = 0.5$$

Mức độ ảnh hưởng của biểu đồ có thể tính từ trục tung. Chú ý biểu đồ có 8 đường ngang (bắt đầu từ 0 đến 7), nhưng chúng ta chú ý từ 1 đến 6. Do đó, mức độ ảnh hưởng trên biểu đồ là:

$$ESgraph = (6 - 1) / 1 = 5$$

Từ đó, chúng ta có thể ước tính yếu tố dối là:

$$LF = 5 / 0.5 - 10.$$

Biểu đồ này có yếu tố dối quá cao. Chính yếu tố này giải thích tại sao chúng ta cảm nhận rằng mức độ ảnh hưởng rất cao, nhưng trong thực tế thì không hẳn như thế.

Giới báo chí thường hay dùng biểu đồ một cách … dối gian. Một ví dụ khác về yếu tố dối có thể xem qua biểu đồ dưới đây (trích từ một cuốn sách của Tufte). Biểu đồ cho thấy năm 1978, mỗi gallon xăng chạy được 18 mile, nhưng đến năm 1985 thì mỗi gallon xăng chạy được 25 mile, tức là xăng dầu càng ngày càng có hiệu suất kinh tế hơn.

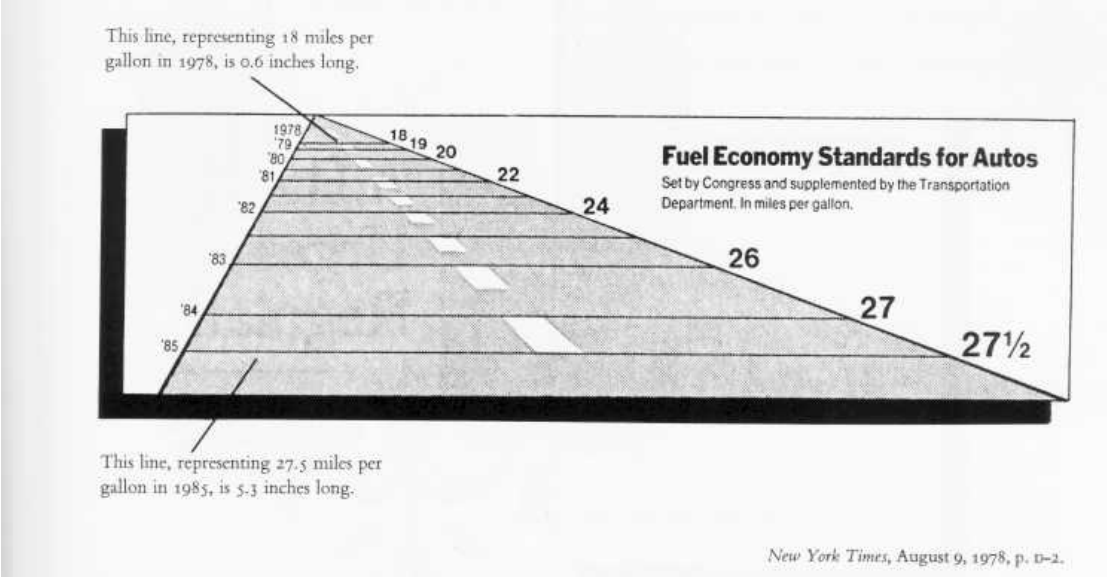
Nhưng vì cách trình bày biểu đồ thiếu thành thật, nên làm cho chúng ta có ấn tượng rất tốt. Nhưng nhìn kĩ thì số liệu của năm 1978 là 18 inch dài, còn năm 1985 là 27.5 inch. Mức độ ảnh hưởng thật sự (tức từ dữ liệu) là:

$$ESdata = (27.5 - 18) / 18 = 0.53 .$$

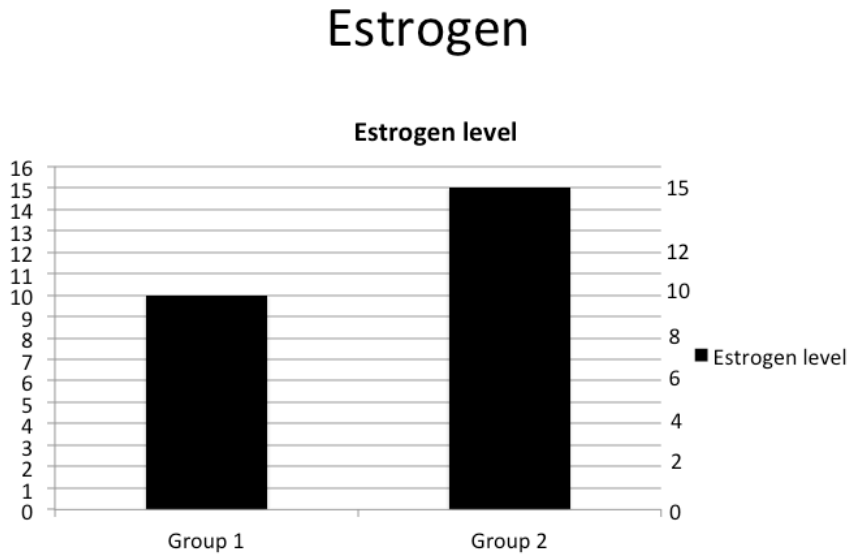
Nhưng mức độ ảnh hưởng qua cách thiết kế biểu đồ thì rất cao. Chú ý rằng trục hoành cho năm 1978 là 0.6 inch, còn năm 1985 là 5.3 inch. Do đó, mức độ ảnh hưởng ảo trên biểu đồ là:

$$ESgraph = (5.3 - 0.6) / 0.6 = 7.83.$$

Như vậy, yếu tố dối gian là gần bằng 15! (LF = 7.83 / 0.53 = 14.8).



Nhưng biểu đồ dưới đây thì không có yếu tố gian dối, vì yếu tố dối bằng 1. (Các bạn có thể tính để kiểm tra).



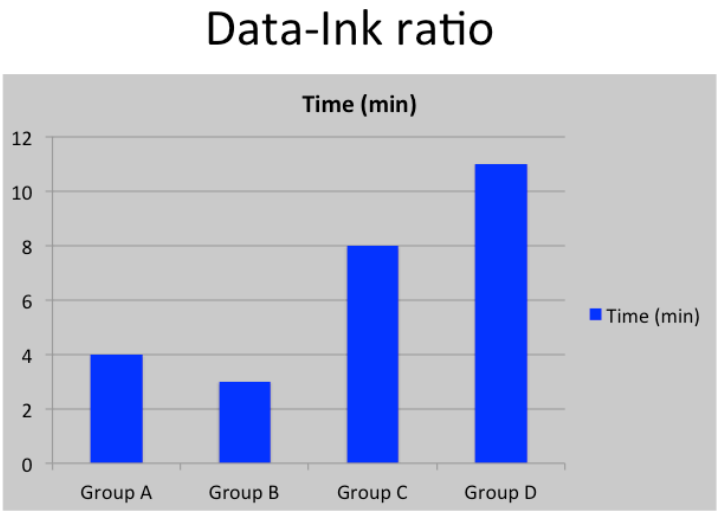
Tỉ số dữ liệu trên mực in (Data-ink ratio)

Một nguyên tắc quan trọng trong thiết kế biểu đồ là sử dụng mực in để trình bày dữ liệu chứ không phải để trang trí cho biểu đồ. Do đó, Tufte đề nghị dùng tỉ số mực in dành cho dữ liệu trên tổng số lượng mực in để đánh giá một biểu đồ. Nói

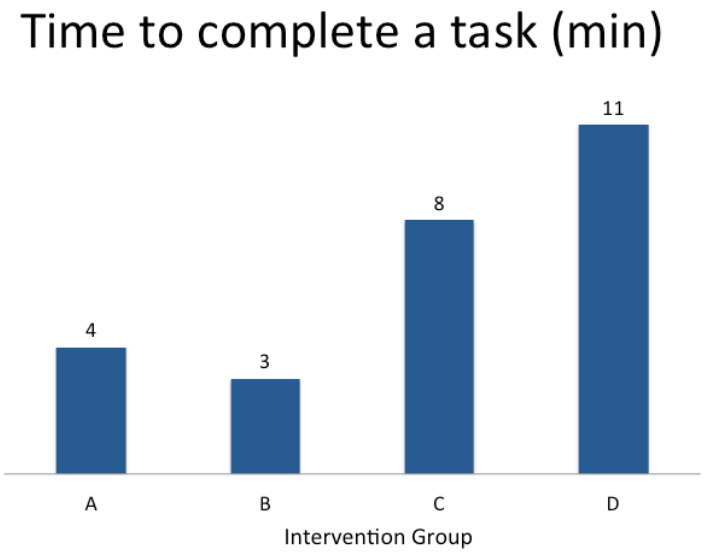
cách khác, gọi DIR (data-ink ratio) là tỉ số dữ liệu và mực in:

DIR = (số mực dùng cho dữ liệu) / (tổng số mực dùng trong biểu đồ)

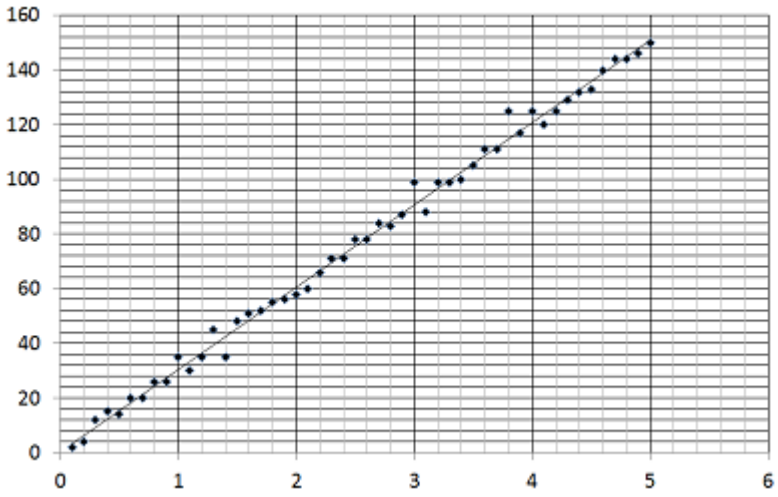
Tỉ số này cũng nên gần bằng 1. Tỉ số này cũng có thể hiểu như là tỉ số của tín hiệu trên nhiễu (signal over noise ratio). Theo đó, nên xoá bỏ những mực in không dùng cho dữ liệu hay thừa thải. Để minh hoạ cho khái niệm DIR, chúng ta có thể xem qua biểu đồ dưới đây:



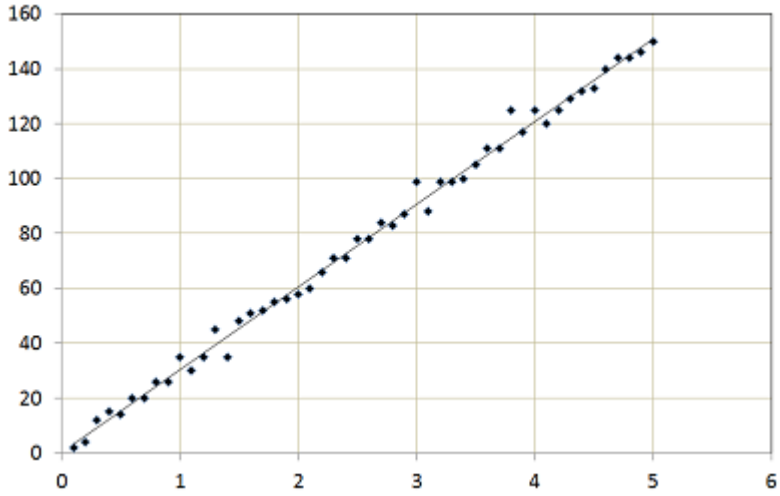
Trong biểu đồ trên, chúng ta dễ dàng thấy có quá nhiều mực dành cho trang trí. Thứ nhất là màu nền (màu xám nhạt) là không cần thiết. Thứ hai là *legend*, Time (min), cũng không cần thiết. Thứ ba là những đường ngang cũng không cần thiết. Ngay cả cách viết "Group A", "Group B", v.v. lặp lại chữ "Group" đến 4 lần! Biểu đồ trên có thể thiết kế lại như sau. Ngay cả cách thiết kế này cũng chưa tối ưu, nhưng có thể chấp nhận được.



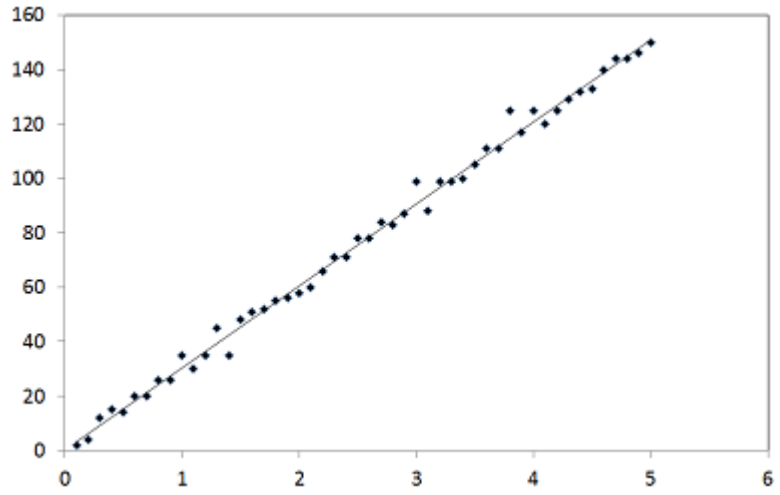
Dưới đây là một ví dụ về biểu đồ có quá nhiều mực cho trang trí:



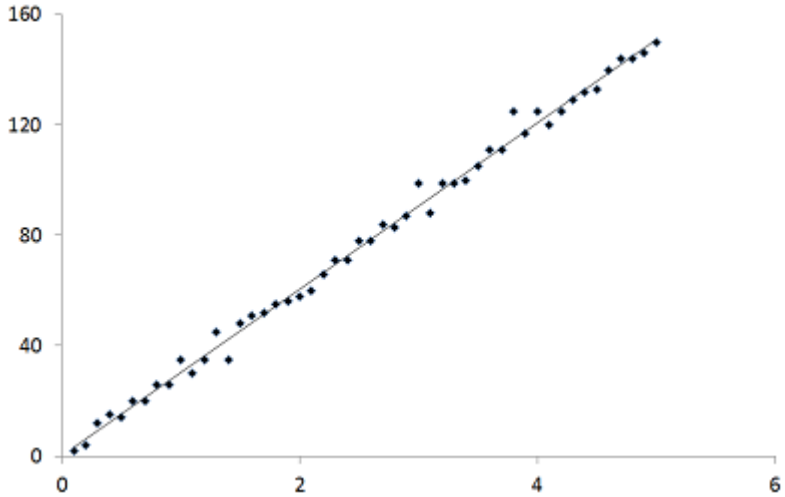
Biểu đồ này có quá nhiều *gridlines* dễ làm cho người xem mất tập trung. Có thể đơn giản thành:



Thật ra, nếu mục tiêu là chỉ ra mối tương quan thì những đường ngang đó cũng không cần thiết, và biểu đồ có thể đơn giản thành:



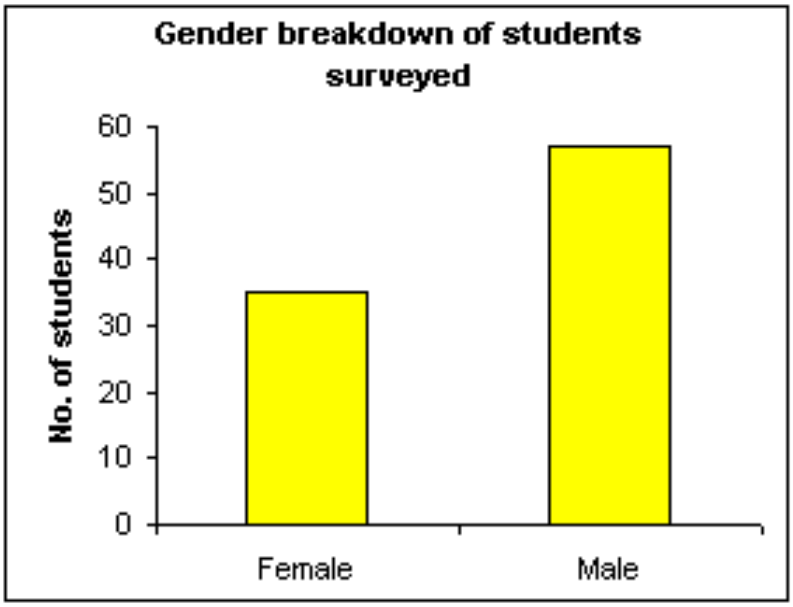
Ngay cả những đường ranh cũng không cần. Do đó, biểu đồ có thể cải tiến thành:



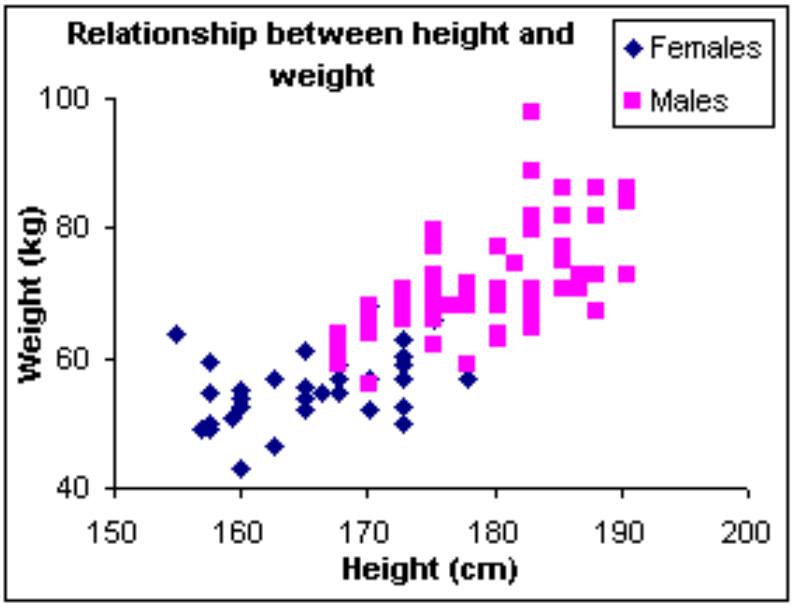
Mật độ dữ liệu

Tufte định nghĩa mật độ dữ liệu (data density index hay DDI) là số số liệu tính trên diện tích của biểu đồ. Nguyên tắc chung là tối đa hoá DDI, vì mục tiêu chính của nhà khoa học là trình bày dữ liệu càng nhiều càng tốt.

Biểu đồ dưới đây trình bày số đối tượng nghiên cứu cho nhóm nam và nữ. Trong thực tế, biểu đồ này rất vô dụng vì tất cả chỉ có 2 số liệu mà thôi, nhưng chiếm rất nhiều không gian. Nếu chúng ta đo chiều cao và chiều ngang của biểu đồ (có thể tính bằng cm) thì sẽ có diện tích. Nhưng giả dụ như diện tích của biểu đồ là 10 cm^2, thì mật độ dữ liệu chỉ 2 / 10 = 0.2, tức rất thấp. Trong trường hợp này, tác giả không cần đến biểu đồ, mà chỉ đơn giản mô tả bằng chữ là đủ.



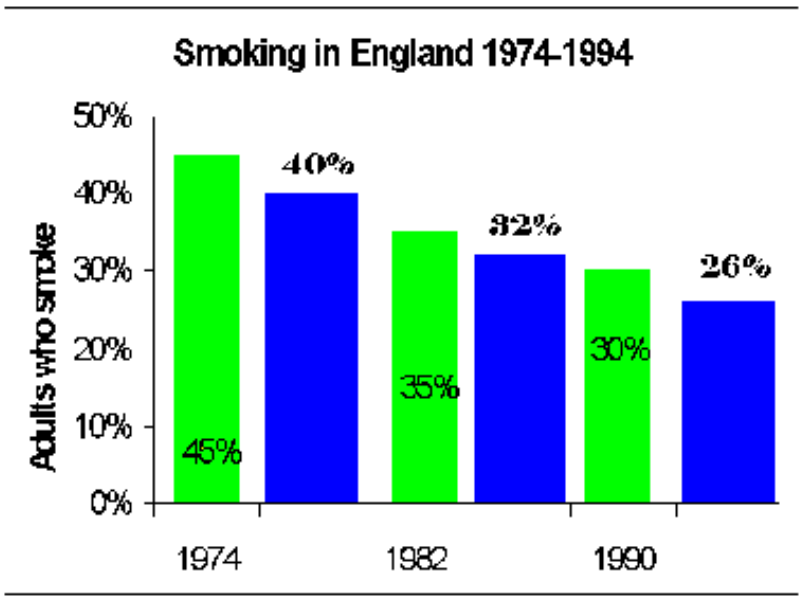
Biểu đồ dưới đây thể hiện mối tương quan giữa chiều cao (trục hoành) và trọng lượng (trục tung). Tác giả còn dùng màu để phân biệt dữ liệu cho nam và nữ. Biểu đồ có rất nhiều dữ liệu và thông tin. Đây là biểu đồ có mật độ dữ liệu cao, và có thể xem là rất tốt.



Edward Tufte làm một nghiên cứu nhỏ để so sánh mật độ dữ liệu của các tạp chí khoa học phổ thông và khoa học chuyên môn. Kết quả cho thấy tập san khoa học như *Nature* có mật độ dữ liệu cao nhất (7.4) so với *Scientific American* (0.8) và *Times* (2.8). Bài học ở đây là để tăng cao xác suất công bố trên những tập san lớn, cần chú ý đến tối ưu hoá mật độ dữ liệu trong biểu đồ.

Nhất quán trong cách thể hiện dữ liệu

Một nguyên tắc quan trọng khác trong thể hiện dữ liệu là trình bày những biến đổi của dữ liệu, chứ không phải thay đổi hình thức (như màu) để thể hiện một dữ liệu. Biểu đồ dưới đây là một ví dụ cho sự "vi phạm" nguyên tắc vừa đề cập:

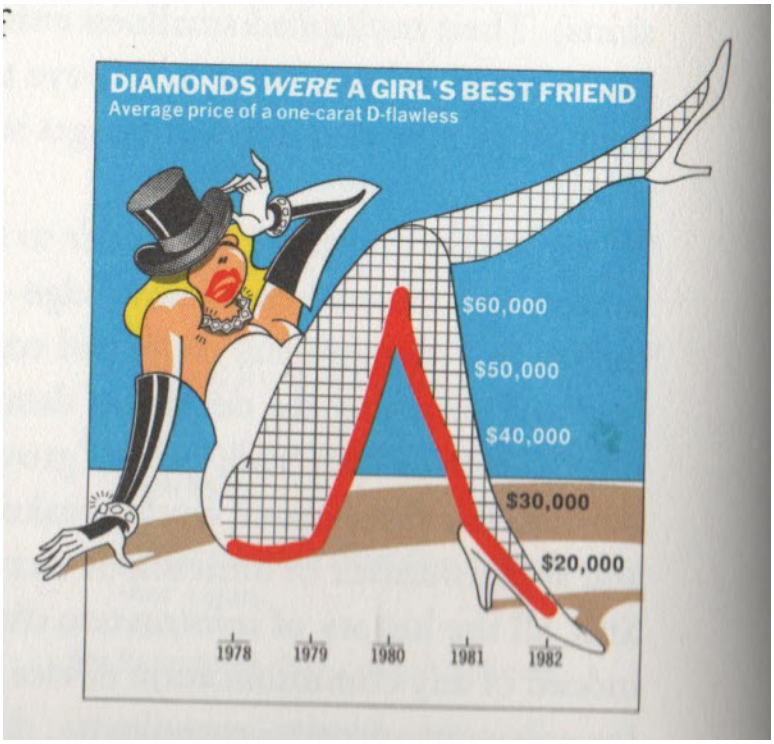


Tác giả dùng hai màu xanh một cách luân phiên để chỉ mô tả tỉ lệ hút thuốc ở Anh. Năm thì không rõ ràng, đáng lẽ phải là 1974, 1978, 1982, ..., 1994, nhưng tác giả để cho người đọc phải suy đoán. Đó là một điều đại kị trong phương pháp trình bày dữ liệu. Ngoài ra, những con số phần trăm (45%, 40%, v.v.) hình như được đặt vào những vị trí rất tuỳ tiện, chứ chẳng theo một qui luật nào cả. Có thể xem đây là một biểu đồ rất kém. Biểu đồ này có thể thiết kế lại tốt hơn, nhưng tôi để cho bạn đọc suy nghĩ và thử làm :-).

Tránh biểu đồ rác rưởi (Chart junk)

Thuật ngữ "Chart Junk" cũng là một sáng kiến của Edward Tufte. Ông gọi biểu đồ rác rưởi là cách thể hiện dữ liệu một cách "hoa hoè" hay loé loẹt. Đây là cách thể hiện dữ liệu của giới báo chí hay nghệ sĩ. Những người này vì không am hiểu khoa học, nên hay lạm dụng những hình ảnh làm độc giả thiếu tập trung vào thông điệp chính của dữ liệu. Cần tránh những biểu đồ rác rưởi.

Một ví dụ tiêu biểu về biểu đồ rác rưởi mà Edward Tufte lấy ra để làm minh hoạ là biểu đồ dưới đây. Biểu đồ trình bày giá của kim cương từ năm 1978 đến 1982. Thay vì đường biểu diễn màu đỏ là đủ, người thiết kế biểu đồ cho thêm hình ảnh một cô gái trong tư thế gợi cảm. Với biểu đồ này, chắc chắn làm cho phần lớn người đọc không chú ý vào dữ liệu mà chăm chú nhìn vào cô gái, và thế là tác giả không đạt được mục tiêu của mình.



Biểu đồ có thể giúp cho chúng ta "dấn thân" vào chủ đề nghiên cứu mà có khi chữ không làm được. Thiết kế biểu đồ tốt cũng đòi hỏi nỗ lực cao như viết một bài báo khoa học. Một biểu đồ tốt có thể đi vào lịch sử và tồn tạo với thời gian rất lâu. Chúng ta hay thấy có nhiều sách giáo khoa hay những bài giảng có những biểu đồ thuộc vào hạng cổ điển, vì những biểu đồ đó chuyển tải thông tin đầy đủ và đạt những chuẩn mực về thiết kế biểu đồ mà tôi trình bày trên đây. Do đó, cần phải đầu tư thời gian và công sức vào cách trình bày dữ liệu và thiết kế biểu đồ.

Trước khi soạn một biểu đồ, cần phải trả lời những câu hỏi sau đây:

- Ai là độc giả của biểu đồ, hay ai sẽ dùng?
- Chọn hình thức thể hiện (biểu đồ thanh, biểu đồ tán xạ, v.v.)
- Sắp xếp dữ liệu thích hợp cho trục tung và trục hoành.
- Thêm vào các biến cần thiết.
- Biên tập nhiều lần để tăng mật độ dữ liệu.

Sau đó là tuân thủ theo 4 nguyên tắc vừa mô tả trên. Xin nhắc lại đó là nguyên tắc tối ưu hoá yếu tố dối, tỉ số dữ liệu trên mực in, tỉ số dữ liệu trên diện tích biểu đồ, và tránh những hình thức màu mè (rác rưởi) dễ làm cho người đọc xa rời thông điệp chính của số liệu. Hi vọng rằng những nguyên tắc và chỉ dẫn trên đây sẽ

giúp cho các bạn có được một bài báo khoa học tốt và những biểu đồ đi vào lịch sử.