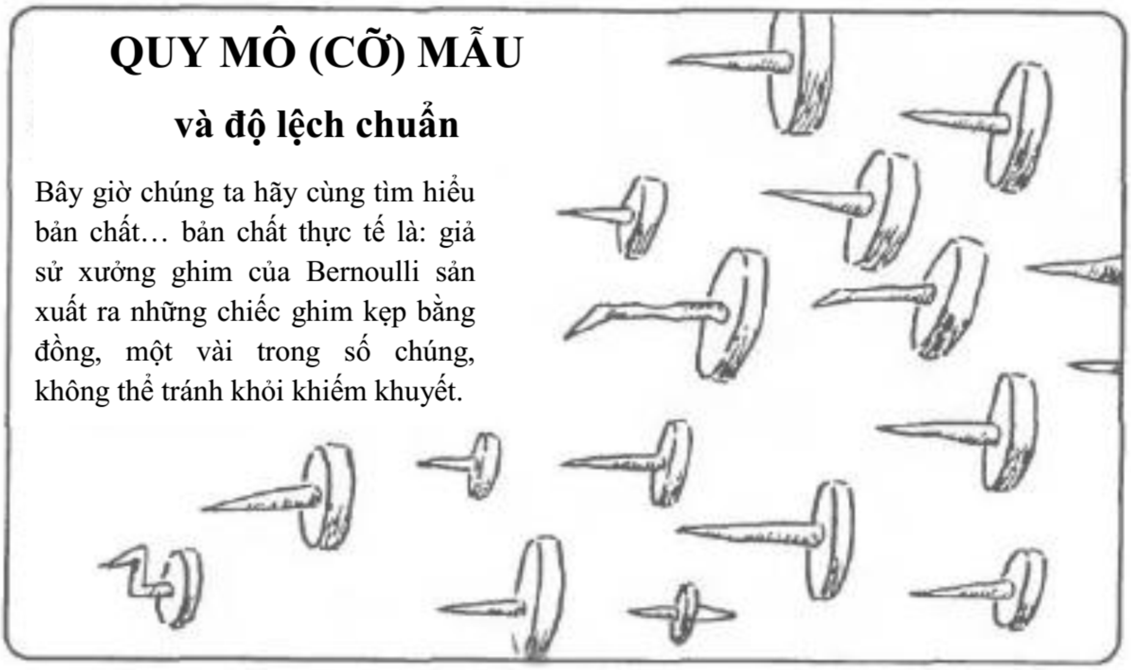
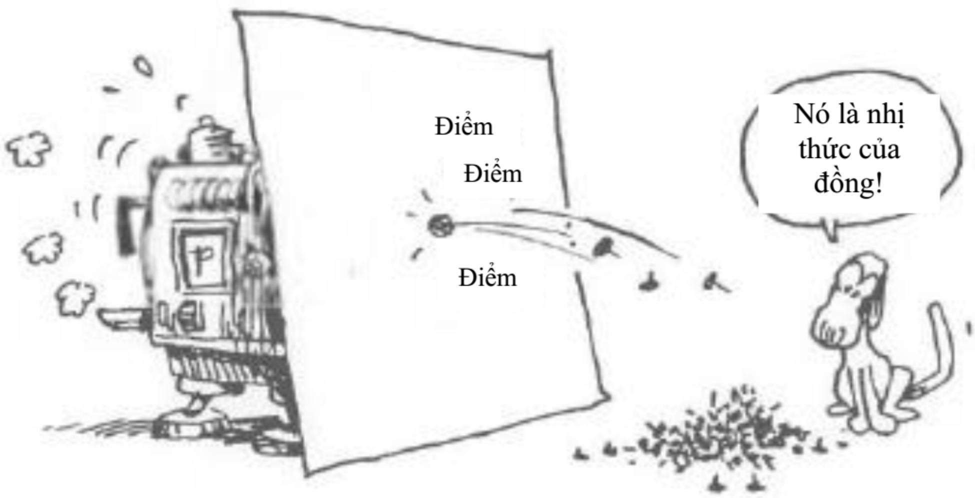
CHƯƠNG 6: CHỌN MẪU

(tiếp theo)



Bạn đọc khôn ngoan sẽ phát hiện ra điều này giống như một hệ thống Bernoulli: Mỗi chiếc ghim mới là kết quả của thử nghiệm Bernoulli với xác suất thành công là p (chẳng hạn không có khuyết tật) và (1-p) là xác suất thất bại (ví dụ có khuyết tật).



Chúng ta nghĩ về tình huống này như thể nếu có một điều tiềm ẩn nhưng hiển nhiên, "ảnh hưởng Bernoulli" chi phối các kết quả chúng ta quan sát theo cái gọi là "thế giới thực".

Bởi vì ảnh hưởng Bernoulli là vô hình, nên không biết p bằng bao nhiêu, nhưng chúng ta lại muốn tìm kiếm. Bởi vậy chúng ta chọn một mẫu ngẫu nhiên của n chiếc ghim, và tìm thấy x chiếc không bị lỗi.

Bây giờ, tỉ lệ thành công trong mẫu có thể là một giá trị nào đó xung quanh p... bởi vậy chúng tôi gọi nó là $\hat{p}$… đọc là “p mũ”.

$$\hat{p} = \frac{x}{n}$$

$\hat{p}$ là tỉ lệ lượng thành công x chia cho cỡ mẫu n. Ví dụ, nếu p là 0.85 và chúng ta lấy mẫu n = 1000 ghim, có thể chúng ta tìm được x = 832 chiếc tốt, $\hat{p}$ bằng 0.832.

Chúng ta thắc mắc: ước lượng này tốt như thế nào?

Và chúng ta trả lời bởi một câu hỏi khác: câu hỏi đầu tiên có ý nghĩa gì?

Chúng ta không thể biết sự chính xác khác nhau giữa $\hat{p}$ và p, bởi vì chúng ta không biết giá trị của p. Câu hỏi thực tế này là: nếu chúng ta lấy nhiều mẫu của 1000 chiếc ghim và quan sát $\hat{p}$ của mỗi mẫu, thì những giá trị này của $\hat{p}$ phân bố xung quanh p như thế nào?

Thực tế, những giá trị $\hat{p}$ này rất giống một biến ngẫu nhiên: sự lựa chọn của n đơn vị mẫu là một thí nghiệm ngẫu nhiên, và quan sát $\hat{p}$ là một kết quả bằng số!

Để chính xác, nếu X là số thành công trong mẫu, thì X chính là biến nhị thức ngẫu nhiên (n lần thử, xác suất p)… chúng ta xác định được tỉ lệ quan sát là một biến ngẫu nhiên:

$$\hat{P} = \frac{X}{n}$$

Sau khi hiểu tất cả về X, chúng ta kết luận ngay được một vài yếu tố về $\hat{P}$

1) Trung bình của $\hat{P}$ là $E[\hat{P}] = p$

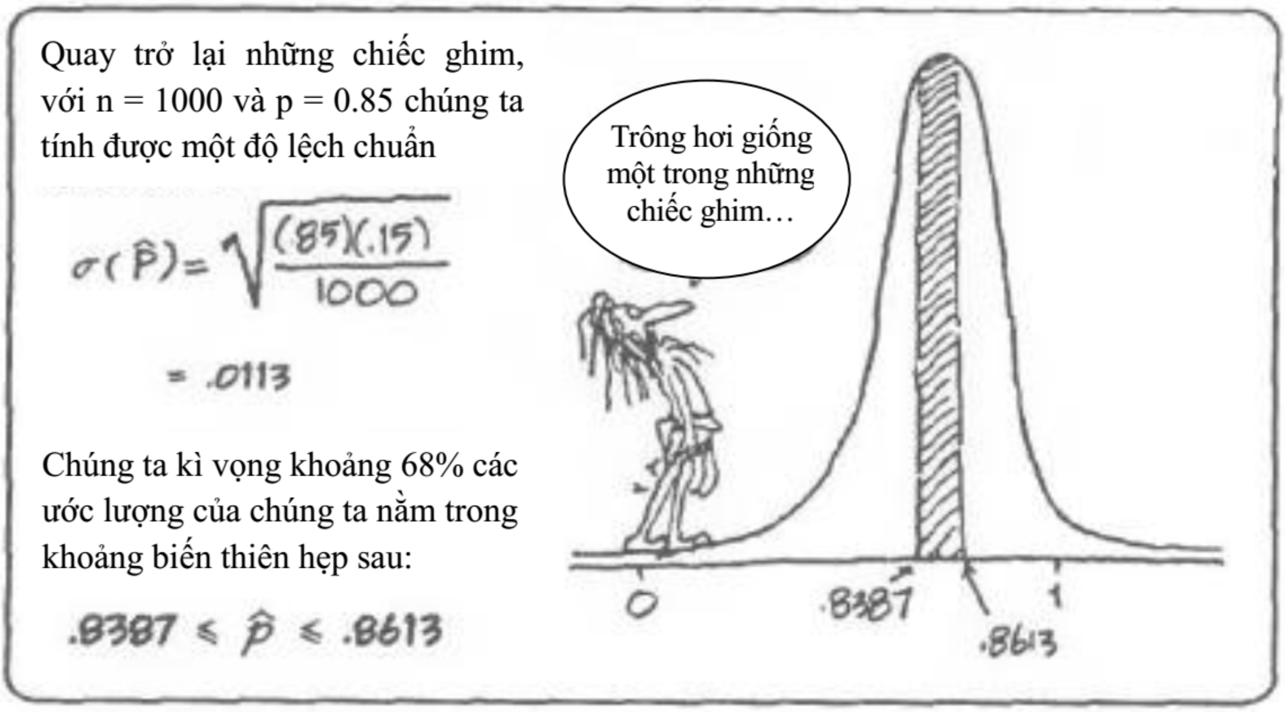
2) Độ lệch chuẩn của $\hat{P}$ là

$$\sigma(\hat{P}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

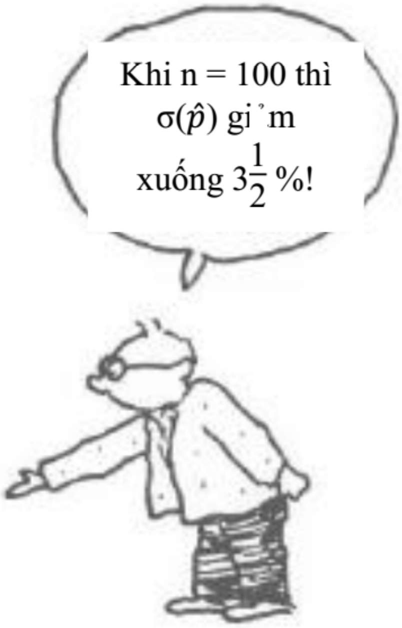
3) n lớn thì $\hat{P}$ gần phân bố chuẩn.

Bạn đã có tất cả rồi đó! giá trị quan sát $\hat{P}$ sẽ là trung tâm của p (đừng ngạc nhiên), còn độ lệch chuẩn của chúng, hoặc khoảng mở rộng, chính là tỉ lệ của những con số thần kì mà chúng ta đã đề cập đến ở phần đầu của chương:

Và, cũng bởi $\hat{P}$ gần như chuẩn, chúng ta có thể sử dụng quy tắc ngón tay cái để kết luận rằng khoảng 68% tất cả các ước lượng sẽ nằm trong độ lệch chuẩn của giá trị p đúng.



Độ lệch chuẩn của $\hat{P}$ là một thước đo của **sai số mẫu**. Như chúng ta đã thấy, nhị thức β, sai số mẫu này tỷ lệ nghịch với $\sqrt{n}$. Khi tăng cỡ mẫu 4 lần thì $\sigma(\hat{P})$ cũng giảm đi 2 lần.



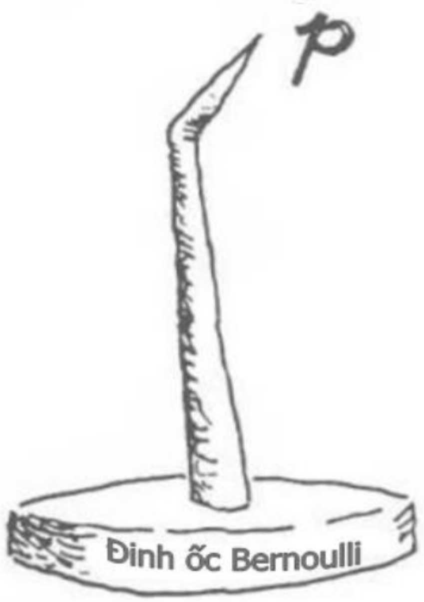
Cỡ mẫu cho những chiếc ghim, p = 0.85

n	1	4	16	25	100	10000
$\sqrt{n}$	1	2	4	5	10	100
$\sigma(\hat{P})$	0.357	0.1785	0.089	0.071	0.357	0.0036

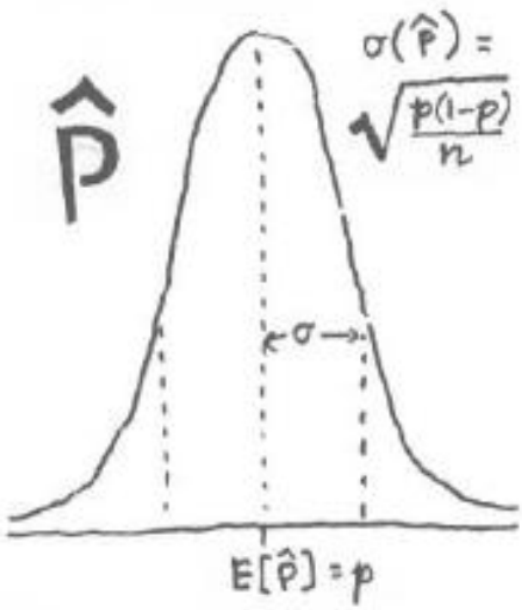
Chú ý: một ước lượng là một đo lường hoặc một quan sát đơn. Hàm ước lượng chính là thước đo của việc ước lượng. Trong trường hợp này, công thức ước lượng chính là một biến ngẫu nhiên $\hat{P} = \frac{X}{n}$.

Hầu hết các số liệu thống kê đều liên quan đến 4 bước của quá trình mà chúng ta cần trải qua như sau:

Xác định tổng thể khi không biết tham số



Tìm công thức ước lượng, phân bổ mẫu theo lý thuyết và độ lệch chuẩn.

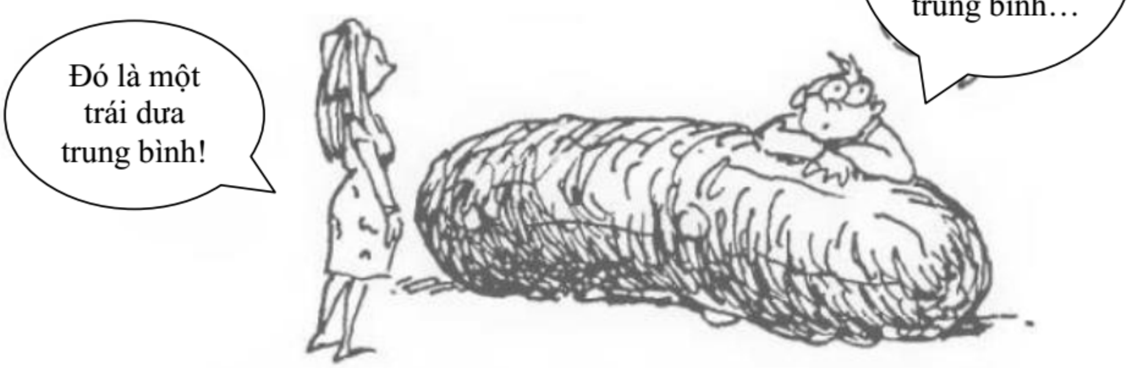


Chọn mẫu ngẫu nhiên thực tế và tìm ước lượng.

Báo cáo kết quả và sai số thống kê hoặc sai số mẫu của nó

Phân phối TRUNG BÌNH mẫu

Bây giờ chúng ta chuyển từ những chiếc ghim bằng đồng sang những trái dưa muối…



Những người sản xuất dưa muối ở California muốn biết độ dài trung bình của mỗi trái dưa mà không cần kiểm tra từng trái. Họ chọn ngẫu nhiên n trái dưa và đo độ dài của chúng được $x_1, x_2, \ldots, x_n$.

Bây giờ bạn đã quen với ý tưởng mỗi X_i là một biến ngẫu nhiên: kết quả bằng số của một biến ngẫu nhiên.

Nếu μ là độ dài trung bình của các trái dưa, và σ là độ lệch chuẩn của phân bố các độ dài, thì

$$E[X_i] = \mu$$
$$\sigma(X_i) = \sigma$$

Với mỗi i (vì x_i có thể là độ dài của một số trái dưa).

Bây giờ chúng ta hãy nhìn vào trung bình mẫu: Độ dài trung bình của các trái dưa đã chọn là một biến ngẫu nhiên được cho bởi:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Như trước đó chúng tôi muốn biết trung bình mẫu này "gần" với μ như thế nào, nếu những mẫu này được chọn nhiều lần thì phân bố của $\bar{X}$ là gì? bởi chúng ta đã biết X_1, X_2, …, X_n. Chúng ta cũng biết rằng

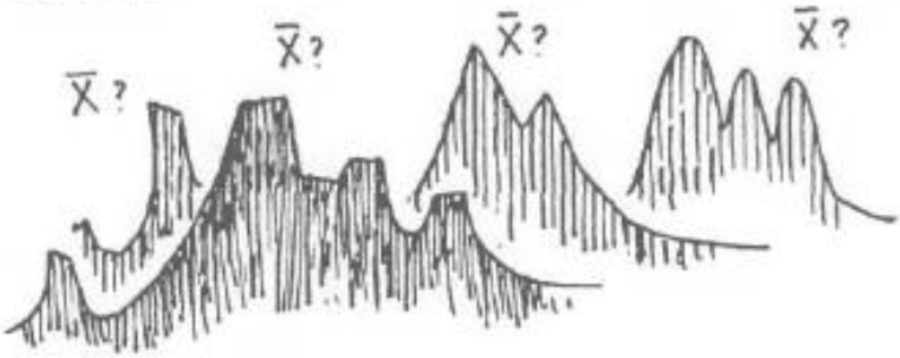
$$E[\bar{X}] = \mu$$

$$\sigma(\bar{X}) = \sigma/\sqrt{n}$$

Một lần nữa, chúng ta lại thấy mẫu số thần kỳ xuất hiện trong công thức khoảng mở rộng của trung bình các mẫu là

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

Nhưng chúng ta không biết hình dạng phân bố của $\bar{X}$. Phân bố xác suất mẫu $\hat{p}$ hầu hết là phân bố chuẩn, vì nó được dựa trên một biến ngẫu nhiên nhị thức. Còn $\bar{X}$ thì sao? Công thức ước lượng trung bình mẫu là gì???



(Còn nữa)

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