

Giới thiệu thống kê sinh học

Nguyễn Quang Vinh

Giới thiệu

- Thống kê là **khoa học** nghiên cứu **dữ liệu**, một ngành học về **sự bất định**
- **Sinh thống kê**: dữ liệu từ y khoa, giáo dục, tâm lý, nông nghiệp, kinh tế, thương mại...
- **Xã hội hiện đại**:
 - *Đọc, Viết*
 - *Suy nghĩ mang tính thống kê*: Để đưa ra các kết luận **mạnh nhất** có thể có được từ nguồn **dữ liệu hạn chế**.

Thống kê

Thống kê mô tả

- Đo lường xu hướng **tập trung**
- Đo lường **phân tán**
- Đo lường **vị trí** dữ liệu
- Đo lường **hình dạng** phân phối

Thống kê suy lý

- Ước lượng
- Kiểm định giả thuyết → đưa ra quyết định
- + Thống kê **tham số**
- + Thống kê **phi tham số** << thống kê **phân phối bất kỳ**

Thống kê suy lý

Ước lượng

Tại sao cần ước lượng?

- Tổng thể **vô hạn** → không thể khảo sát hết
- Tổng thể **hữu hạn** → chi phí, thời gian.
- Ngoài ra, ước lượng có thể **sớm đưa ra** kết luận, không chờ đến khi quan sát toàn bộ tổng thể

Ước lượng gì?: điểm & khoảng tin cậy (rất hữu ích vì cho biết khả năng dao động của giá trị cần nghiên cứu).

KHOẢNG TIN CẬY CỦA TRUNG BÌNH TỔNG THỂ

Nói chung, ước lượng khoảng có công thức

Giá trị ước lượng điểm \pm (hệ số tin cậy) \times (sai số chuẩn)

Khi mẫu được chọn từ tổng thể có phân phối *bình thường* với *phương sai biết trước*, ước lượng khoảng cho trung bình μ sẽ là:

$$\bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}}$$

Cách diễn giải kết quả khoảng ước lượng theo công thức này

- Nếu lấy mẫu lặp đi lặp lại càng nhiều lần, từ tổng thể có phân phối bình thường, $100(1 - \alpha)\%$ của tất cả các khoảng ước lượng tính theo công thức trên sẽ chứa trung bình của tổng thể μ
- Con số $(1 - \alpha)$, gọi là *hệ số tin cậy*, &
Khoảng $\bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}}$, gọi là *khoảng tin cậy của μ*

Cách diễn giải thực tế

- Chúng ta tin cậy ở mức $100(1 - \alpha)\%$ là khoảng ước lượng tính được này

$$\bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}}$$

sẽ chứa trung bình của tổng thể, μ

- $E =$ phạm vi sai số = sai số tối đa = sai số chấp nhận được trên thực tế / lâm sàng :

$$E = z_{\alpha/2} \sigma_{\bar{x}} = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Các giá trị sử dụng nhiều nhất:

.90, .95, .99, liên hệ với hệ số tin cậy, lần lượt là:
1.645, 1.96, 2.58

$$P(-1.96 \leq z \leq 1.96) = .95$$

CHỌN MẪU TỪ TỔNG THỂ CÓ PHÂN PHỐI NONNORMAL

→ Việc chọn mẫu từ:

- Tổng thể có phân phối nonnormal
- Tổng thể có hình dạng không biết trước

→ Lấy cỡ mẫu đủ lớn → *định lý giới hạn trung tâm*

TÍNH CỠ MẪU CHO ƯỚC TÍNH TRUNG BÌNH

$$E = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$\Leftrightarrow n = \left[Z_{\alpha/2} \frac{\sigma}{E} \right]^2 \quad (1)$$

(1) công thức tính cỡ mẫu chung

Ghi chú

- * Thông thường không biết phương sai σ^2 → cần phải ước lượng σ^2
- * Việc ước lượng σ^2 từ các nguồn sau đây:
 1. Mẫu nghiên cứu thử
 2. Kết quả nghiên cứu trước hoặc tương tự
 3. $\sigma \approx R/4$ (hoặc $R/6$) (phân phối xấp xỉ bình thường)
 4. $s \approx \text{IQR}/1.35$

Thống kê suy lý

Kiểm định giả thuyết → đi đến một quyết định

- Giúp nghiên cứu viên *đưa ra một quyết định* liên quan đến tổng thể bằng cách *khảo sát một mẫu* lấy ra từ tổng thể đó
- Giả thuyết (GT): một phát biểu liên quan đến một hoặc nhiều tổng thể
- Hai loại giả thuyết:

(1) *Giả thuyết nghiên cứu:*

- Là kết quả của nhiều năm quan sát
- Trực tiếp dẫn đến GT thống kê

(2) *Giả thuyết thống kê:*

Là GT được phát biểu theo cách thức có thể đánh giá bằng các kỹ thuật kiểm định thống kê phù hợp: H_0 & H_A

<p>Các trường hợp mắc sai lầm loại I & sai lầm loại II (4 khả năng)</p>		<p>Hiện trạng thực sự (sự thật trong tổng thể)</p>	
		<p><i>Có mối liên hệ giữa tiếp xúc & kết cục (H₀ sai)</i></p>	<p><i>Không có mối liên hệ giữa tiếp xúc & kết cục (H₀ đúng)</i></p>
<p>Kết quả của nghiên cứu từ mẫu → Kết luận:</p>	<p>Bác bỏ H₀</p>	<p>Quyết định đúng</p>	<p>Sai lầm loại I</p>
	<p>Không bác bỏ H₀</p>	<p>Sai lầm loại II</p>	<p>Quyết định đúng</p>

Giá trị $p^{(*)}$

- Xác suất thấy được sự khác biệt nếu giả thuyết H_0 đúng (+++).

(Xác suất có được kết quả nghiên cứu do tình cờ nếu giả thuyết H_0 đúng).

- Cách diễn tả sự tin tưởng vào H_0 .
- Mốc (ngưỡng) quyết định.

() giá trị “mức độ ngạc nhiên”*

Giá trị p

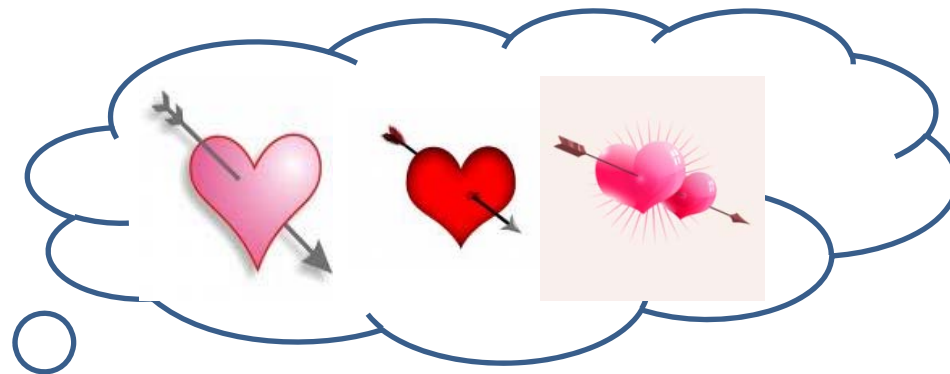
- Bác bỏ giả thuyết H_0 khi giá trị $p < \alpha$
- Ngưỡng ý nghĩa thống kê **không** đồng nghĩa với tình huống *“được ăn cả, ngã về không”* (quá lệ thuộc vào giá trị p).
 - Cách giải thích khi giá trị $p > \alpha$

Kiểm định giả thuyết

Một phía so với Hai phía

Lực của một phép kiểm thống kê

Ước lượng & Kiểm định giả thuyết *khoảng tin cậy & giá trị p*



Mục tiêu của người làm lâm sàng

- **Hiểu được** phần thống kê trong hầu hết các bài báo khoa học trong các tạp chí y khoa.
- **Tránh** bị rối trí bởi các loại thông kê vô nghĩa.
- Tự thực hiện các **phép tính thống kê đơn giản**.
- Biết cách sử dụng các **chương trình thống kê đơn giản** để xử lý dữ liệu.
- **Có khả năng tham khảo** các thống kê nâng cao hơn hoặc trao đổi với các chuyên gia thống kê (mà không cần phiên dịch).

Hai vấn đề

- Sự khác biệt quan trọng **thường bị che khuất** (biến thiên sinh học và/hoặc sự không chuẩn xác trong thực nghiệm)
- **Khái quát hóa kết quả quá mức**

Làm sao để vượt qua

- Phán đoán có căn cứ dựa trên lâm sàng & khoa học
- Trực quan
- Trung thực

Thống kê khiến cho người làm
nghiên cứu trở nên

*người giải quyết vấn đề một
cách độc lập & thấu đáo*

Ứng dụng cho xử lý dữ liệu

Rất quan trọng!

Sắp đặt xử lý đúng?

→ *Chuẩn bị kỹ lưỡng các bảng nội dung xử lý trống*

Bài tập

1. Khảo sát 100 ca sanh trong một dân số, có trung bình của trẻ sơ sinh là 3.050 g; 95% khoảng tin cậy (2.950 - 3.150 g).

Hãy giải thích các kết quả trên.

2. Cách giải thích ngắn gọn về giá trị p:

-

-

-

Tài liệu tham khảo

1. Intuitive Biostatistics. Harvey Motulsky. Oxford University Press, 2010.
2. Business Statistics Textbook. Alan H. Kvanli, Robert J. Pavur, C. Stephen Guynes. University of North Texas, 2000.
3. Biostatistics: A Foundation for Analysis in the Health Sciences. Wayne W. Daniel. Georgia State University, 1991.