# XÁC XUẤT - THỐNG KÊ VÀ CÁC ỨNG DỤNG TRONG NGHIÊN CỨU Y HỌC

*Nguyễn Công Bình* •

**Tóm tắt**: Xác xuất - Thống kê là môn khoa học đo lường sự may rủi về các hiện tượng trong đời sống kinh tế, xã hội, từ đó giúp chúng ta có được quyết định đúng đắn, hợp lý với rủi ro thấp nhất. Bài viết này làm rõ ý nghĩa, tác dụng của môn khoa học Xác suất - Thống kê đối với lĩnh vực y học và đối với giảng dạy sinh viên thuộc khối sức khỏe tại trường đại học.

**Từ khóa**: xác xuất, thống kê, y học, rủi ro.

***Summary***: *Probability - Statistics is the science of measuring the chance/risk of phenomena in economic and social life, thereby helping us to make the right, reasonable decision with the lowest risk. This article clarifies the meaning and effects of the Science of Probability and Statistics for the medical field and for teaching students at universities health faculties.*

***Keywords***: *probability, statistics, medicine, risk.*

Xác suất – Thống kê, có thể nói ngắn gọn, là môn khoa học đo sự may rủi và giúp ta có quyết định lựa chọn kết luận ít rủi ro nhất. Ở trường Kinh doanh và Công nghệ Hà Nội, Xác suất – Thống kê được giảng dạy cho sinh viên các khoa của khối Sức khỏe. Trong các năm tới, môn học này có thể mở rộng sang dạy cho các chuyên ngành khác của trường. Vậy tại sao Xác suất – Thống kê là môn học không thể thiếu trong nghiên cứu Y – Dược và các ngành khoa học khác?

Trong nghiên cứu Y Dược, chúng ta thường gặp các câu hỏi và các vấn đề sau:

- Liệu một loại vaccine (hay một loại thuốc, một phương pháp chữa bệnh mới) có thực sự tác dụng hay không? Trên thực tế, có trường hợp cả người tiêm và không tiêm vaccine ấy đều mắc bệnh, hoặc cả người tiêm lẫn không tiêm đều không mắc bệnh. Vậy việc dùng vaccine có độc lập với việc mắc bệnh không?

- Tỷ lệ người khỏi bệnh khi dùng loại thuốc mới có thực sự lớn hơn tỷ lệ đó khi dùng loại thuốc sẵn có hay không?

- Có thể ước lượng kích thước của một sản phẩm (hoặc ước lượng tỷ lệ khỏi bệnh khi dùng một phương pháp chữa bệnh mới) nằm trong một khoảng tin cậy với một xác suất tin cậy hay không? Những kết luận rút ra có thể dẫn tới việc bác bỏ một sản phẩm hay một phương pháp chữa bệnh mới?

- Có thể biểu diễn dưới dạng hàm số một đại lượng này theo các đại lượng khác không? Ví dụ, ta muốn biết một quan hệ hàm số giữa Y (dung lượng thở) và X (chiều cao), Z (lồng ngực). Nếu biết được

* Khoa Toán, Trường ĐH KD&CN Hà Nội.

quan hệ này sẽ giúp rất nhiều cho việc phân tích và dự báo. Phương pháp này gọi là *phương pháp hồi quy* trong xác suất.

- Khi gặp hai định lượng có giá trung bình khác nhau, liệu ta có thể kết luận ngay là chúng hơn kém nhau thực sự không? Chẳng hạn, tính chiều cao trung bình của nam thanh niên tuổi từ 18-25 ở hai vùng A, B thấy khác nhau, không thể nói thanh niên vùng A cao hơn (hay thấp hơn) thanh niên vùng B? Kết luận này là nguy hiểm vì số liệu ngay của một vùng đo trong các thời điểm khác nhau cũng khác nhau. Sự khác nhau đó gọi là sai số ngẫu nhiên. Phương pháp so sánh (trung bình, phương sai) lọc bỏ các sai số ngẫu nhiên để chỉ ra sự sai khác đó có phải bản chất không?

Bài toán so sánh trong Xác suất - Thống kê có rất nhiều ứng dụng. Chẳng hạn:

- So sánh trung bình lượng protein trong huyết thanh của bệnh nhi suy dinh dưỡng trước và sau điều trị để xem phương pháp có hiệu quả không?

- So sánh tỷ lệ nhiễm độc chì của một làng nghề và một làng không làm nghề xem có khác nhau và có ý nghĩa không, để từ đó rút ra phương pháp phòng tránh.

Những vấn đề đặt ra ở trên đã được giải quyết bằng những phương pháp, như kiểm định giả thuyết, ước lượng, so sánh, hồi quy,... của môn học Xác suất - Thống kê.

Phải khẳng định rằng, không có kết luận nào trong Xác suất - Thống kê là tuyệt đối đúng, nhưng cũng thật tuyệt vời nếu nhận được một kết luận với độ rủi ro thấp dưới 5%, thậm chí dưới 1%.

Trong các phương pháp của Xác suất - Thống kê, phương pháp kiểm định giả thuyết là bao trùm nhất. Chúng ta cùng tiếp cận thêm về phương pháp này và tìm hiểu một ví dụ ứng dụng nó trong nghiên cứu y học hiện đại.

Khi kiểm định một giả thuyết, ta thường nêu ra một giả thuyết H và một đối thuyết K và tiến hành thu thập số liệu (hoặc dựa trên số liệu có sẵn) rồi tính toán theo các thủ tục thống kê để kết luận nên nhận H hay nhận K. Chẳng hạn, khi nghiên cứu ta có thể gặp các giả thuyết H và đối thuyết K như sau: + H: "Chiều cao X của nam thanh niên tuổi từ 18 đến 25 là đại lượng ngẫu nhiên có phân phối chuẩn" với đối thuyết K là "X không có phân phối chuẩn"; + H: "Việc tiêm phòng một loại bệnh độc lập với việc mắc bệnh tức là tiêm phòng không có tác dụng" với đối thuyết K: "Tiêm phòng có ảnh hưởng (giảm) tỷ lệ mắc bệnh"; + H: "Tỷ lệ người khỏi bệnh điều trị bằng phương pháp mới không khác tỷ lệ khỏi bệnh điều trị bằng phương pháp cũ" và đối thuyết K: "Hai tỷ lệ này khác nhau thực sự"...

Bằng các thủ tục thống kê kiểm định giả thuyết, ta đi đến kết luận nhận H hay K.

Kết luận nào cũng gặp một trong hai loại sai lầm:

- Sai lầm loại 1: Bỏ giả thuyết H nhận đối thuyết K, khi H đúng. Xác suất xảy ra sai lầm này gọi là xác suất sai lầm loại 1;

- Sai lầm loại 2: Nhận giả thuyết H khi H sai, dẫn đến bỏ đối thuyết K. Xác suất của sai lầm này gọi xác suất sai lầm loại 2.

Lý tưởng nhất là tìm được tiêu chuẩn thống kê làm cho cả hai loại xác suất trên đồng thời bằng 0 hoặc đồng thời cực tiểu. Tuy nhiên, người ta cũng chứng minh được rằng, không thể tồn tại tiêu chuẩn thống kê như thế, vì khi xác suất sai lầm này giảm, thì xác suất sai lầm kia lại tăng. Người ta đưa ra giải pháp là, tìm tiêu chuẩn thống kê sao cho xác suất sai lầm loại 1 không vượt qua một ngưỡng $\alpha$ khá nhỏ (thường là 0,05 hay 0,01, thậm chí 0,001, tùy từng yêu cầu của bài toán) và cực tiểu hóa xác suất sai lầm loại 2.

Các bước giải bài toán kiểm định như sau:

+ Từ tập thể gốc X, lấy một mẫu ngẫu nhiên kích thước n: X1, X2, .., Xn , ta thiết lập thống kê F = f(X1, X2,.. , Xn) là một hàm số từ mẫu thu thập với điều kiện phân phối xác suất của F được xác định nếu giả thuyết H là đúng.

+ Vì phân phối của F xác định khi H đúng, nên với giá trị α đã cho, ta luôn tìm được miền ωα sao cho $P(F \in \omega\alpha \mid H) \leq \alpha)$ (tức xác suất để thống kê F nhận giá trị thuộc ωα , nếu giả thiết H là đúng không lớn hơn α). Khi α nhỏ, theo nguyên lý tin tưởng thực hành, thì với một phép thử, kết quả sẽ là F nhận giá trị không thuộc ωα. Nếu chỉ với một lần quan sát có kết quả giá trị của F∈ ωα thì ta có thể nói, giả thuyết H là sai. Kết luận này có thể mắc sai lầm với xác suất α. Giá trị α gọi là mức ý nghĩa của kiểm định và miền ωα gọi là miền bác bỏ giả thuyết H.

Trong mùa dịch SARS-COV2 hiện nay, tất cả các nước đều đang chạy đua tìm vaccine phòng chống bệnh này. Tiếc rằng, đến nay vẫn chưa có vaccine nào được thử nghiệm để áp dụng các kiểm định xác suất - thống kê cho kết luận về nó. Ở đây, xin giới thiệu một mô hình chung để xác định xem một loại vaccine có tác dụng (tức là dùng loại vaccine này có làm thay đổi (giảm) tỷ lệ mắc bệnh) hay không.

Khi điều tra số liệu về dùng một loại vaccine phòng bệnh, ta thu được kết quả tại Bảng 1:

*Bảng1*. **Số liệu điều tra**:

| | Mắc bệnh | Không mắc bệnh | Tổng số |
|---|---|---|---|
| Có tiêm | 18 | 232 | 250 |
| Không tiêm | 92 | 658 | 750 |
| **Cộng** | **110** | **890** | **1.000** |

Đây là số liệu điều tra khi sử dụng một loại vaccine mới để phòng bệnh. Ta thấy có người tiêm vẫn mắc bệnh và cả người không tiêm, nhưng không mắc bệnh. Vậy phải chăng vaccine không có tác dụng? Dựa vào số liệu ta muốn kiểm tra giả thuyết:

+ H: Vaccine không có tác dụng, tức là việc tiêm phòng và mắc bệnh độc lập với nhau.

+ K: Vaccine có tác dụng, tức là việc tiêm phòng có làm giảm số người mắc bệnh.

Ta suy luận như sau:

Nếu H đúng, tức tiêm phòng và mắc bệnh là độc lập, thì tỷ tệ mắc bệnh của người có tiêm và không tiêm phải như nhau và bằng tỷ lệ mắc bệnh của cả cộng đồng.

Tỉ lệ mắc bệnh chung là: $\frac{110}{1000} = 0{,}11$.

Vì vậy, với 250 người có tiêm, số người mắc bệnh phải là: 250 . 0,11 = 27,5 người. Số người có tiêm không bị bệnh là: 250 – 27,5 = 222,5.

Tương tự, ta tính được số người không tiêm mắc bệnh là: 750 . 0,11 = 82,5 và số người không tiêm không bị bệnh là: 750 – 82,5 = 667,5.

Các kết quả này được thể hiện trong Bảng 2 - Bảng số liệu lý thuyết:

*Bảng 2.* **Bảng số liệu lý thuyết**

| | Mắc bệnh | Không mắc bệnh | Tổng số |
|---|---|---|---|
| Có tiêm | 27,5 | 222,5 | 250 |
| Không tiêm | 82,5 | 667,5 | 750 |
| **Cộng** | **110** | **890** | **1.000** |

Ta thấy số liệu giữa hai Bảng thực nghiệm và lý thuyết là khác nhau. Sự khác nhau này có thể là sai số ngẫu nhiên khi điều tra, hoặc do sai lầm khi giả thuyết tỷ lệ mắc bệnh khi tiêm phòng hoặc không tiêm phòng là như nhau?

Do việc điều tra là ngẫu nhiên, nên các số liệu thu được là giá trị của một đại lượng ngẫu nhiên. Ta xét xem tổng bình phương sai lệch các giá trị tương ứng giữa hai bảng giá trị là bao nhiêu. Tổng này được ký hiệu là:

$$x^2 = \frac{(27,5-18)^2}{27,5} + \frac{(92-82,5)^2}{82,5} + \frac{(232-222,5)^2}{222,5} + \frac{(667,5-658)^2}{667,5} = 4,916582$$

Theo lý thuyết Xác suất – Thống kê, đây là giá trị quan sát của một đại lượng ngẫu nhên có phân phối $x^2$ với (2-1) . (2-1) = 1 bậc tự do. Tra bảng phân phối $x^2$ với 1 bậc tự do, ta thấy nó nhận giá trị lớn hơn 3,841 với xác suất 0,05. Ở đây, giá trị tính được là 4,916582 lớn hơn 3,841. Điều này chứng tỏ giả thuyết H của ta bị bác bỏ, tức là ta chấp nhận đối thuyết K: Vaccine thử nghiệm có tác dụng. Kết luận này có khả năng mắc sai lầm 0,05.

Chú ý: Bài toán này cũng có thể giải bằng cách so sánh hai tỷ lệ mắc bệnh của những người tiêm phòng và không tiêm phòng. Kết quả thu được cũng tương tự như kết luận ở trên, tức là tỷ lệ mắc bệnh khi không tiêm phòng lớn hơn hẳn tỷ lệ này khi tiêm phòng. Vậy tiêm phòng loại vaccine thử nghiệm có tác dụng./.

**Ngày nhận bài: 25/05/2020**