

ĐẠI HỌC DUY TÂN
KHOA KHOA HỌC TỰ NHIÊN
BỘ MÔN XÁC SUẤT THỐNG KÊ



BÀI GIẢNG
LÝ THUYẾT XÁC SUẤT VÀ THỐNG KÊ TOÁN
(Lưu hành nội bộ)

Đà Nẵng, Năm 2021

ĐẠI HỌC DUY TÂN
KHOA KHOA HỌC TỰ NHIÊN
BỘ MÔN XÁC SUẤT THỐNG KÊ

ThS. Trần Anh Việt

LÝ THUYẾT XÁC SUẤT VÀ THỐNG KÊ

Khoa: KHOA HỌC TỰ NHIÊN

MỤC LỤC

Trang phụ bìa	i
Mục lục	1
Chương 1 Bản chất của xác suất thống kê	5
1.1 Thống kê mô tả và thống kê suy diễn	6
1.2 Các loại biến và các loại dữ liệu	6
1.2.1 Biến định tính và biến định lượng	6
1.2.2 Các loại thang đo thường được sử dụng	7
1.3 Thu thập dữ liệu và phương pháp lấy mẫu	8
1.3.1 Mẫu ngẫu nhiên (<i>Random Sampling</i>)	9
1.3.2 Mẫu hệ thống (<i>Systematic Sampling</i>)	9
1.3.3 Mẫu phân tầng (<i>Stratified Sampling</i>)	9
1.3.4 Mẫu chùm (<i>Cluster Sampling</i>)	9
1.4 Nghiên cứu quan sát và thực nghiệm	10
1.4.1 Nghiên cứu quan sát	10
1.4.2 Nghiên cứu thực nghiệm	10
1.5 Lợi ích và lạm dụng của thống kê	10
1.6 Máy tính và tính toán	11
Chương 2 Phân bố tần số và đồ thị	15
2.1 Tổ chức dữ liệu	15
2.2 Nhật đồ, đa giác tần số, hình cung	20
2.2.1 Nhật đồ	20
2.2.2 Đa giác tần số	21
2.2.3 Hình cung (<i>Ogive</i>)	21
2.2.4 Biểu đồ tần số tương đối	23
2.2.5 Các dạng phân phối	24
2.3 Một số loại biểu đồ khác	27
2.3.1 Biểu đồ thanh (<i>Bar graph</i>)	27
2.3.2 Biểu đồ Pareto	28
2.3.3 Biểu đồ chuỗi thời gian (<i>Time series graph</i>)	29
2.3.4 Biểu đồ hình tròn (<i>Pie graph</i>)	29
2.3.5 Biểu đồ gây nhầm lẫn	31
2.3.6 Biểu đồ thân và lá (<i>Stem and Leaf Plots</i>)	33

Chương 3	Mô tả dữ liệu	45
3.1	Các đo lường khuynh hướng định tâm (<i>measures of central tendency</i>)	45
3.1.1	Trung bình (<i>mean</i>)	45
3.1.2	Trung vị (<i>median</i>)	48
3.1.3	Mode	48
3.1.4	Trung bình khoảng (<i>midrange</i>)	49
3.2	Các thước đo về độ biến thiên (<i>measures of variation</i>)	51
3.2.1	Khoảng biến thiên (<i>range</i>)	52
3.2.2	Phương sai và độ lệch chuẩn (<i>variance and standard deviation</i>)	52
3.2.3	Hệ số biến thiên (<i>coefficient of variation</i>)	56
3.2.4	Quy tắc tính rợ khoảng (<i>range rule of thumb</i>)	56
3.2.5	Định lý Chebyshev	57
3.2.6	Quy tắc kinh nghiệm (<i>empirical rule</i>)	58
3.3	Các thước đo về vị trí (<i>measures of position</i>)	59
3.3.1	Điểm chuẩn (<i>standard score</i>)	59
3.3.2	Điểm bách phân (<i>percentile</i>)	60
3.3.3	Điểm thập phân và tứ phân (<i>decile and quartile</i>)	62
3.3.4	Giá trị ngoại biên hay cá biệt (<i>outlier</i>)	63
3.4	Phân tích dữ liệu khám phá (<i>exploratory data analysis</i>)	64
Chương 4	Xác suất và các quy tắc đếm	73
4.1	Không gian mẫu và xác suất	73
4.1.1	Các khái niệm cơ bản	73
4.1.2	Xác suất cổ điển	75
4.1.3	Biến cố đối	76
4.1.4	Xác suất thực nghiệm (<i>empirical probability</i>)	78
4.1.5	Luật số lớn	79
4.1.6	Xác suất chủ quan	80
4.2	Quy tắc cộng xác suất	80
4.3	Quy tắc nhân xác suất và xác suất có điều kiện	83
4.3.1	Quy tắc nhân xác suất	83
4.3.2	Xác suất có điều kiện	84
4.4	Các quy tắc đếm	87
4.4.1	Các quy tắc đếm cơ bản	87
4.4.2	Các khái niệm cơ bản về giải tích tổ hợp	88
4.4.3	Xác suất và quy tắc đếm	89
Chương 5	Phân phối xác suất rời rạc	97

5.1	Phân phối xác suất	97
5.2	Trung bình, phương sai, độ lệch chuẩn, kỳ vọng của một phân phối xác suất	99
5.2.1	Trung bình của một phân phối xác suất	99
5.2.2	Phương sai và độ lệch chuẩn của một phân phối xác suất	100
5.2.3	Kỳ vọng của một phân phối xác suất	101
5.3	Phân phối nhị thức	102
5.4	Các loại phân phối khác	103
5.4.1	Phân phối đa thức	103
5.4.2	Phân phối Poisson	104
5.4.3	Phân phối siêu bội (hypergeometric distribution)	105
Chương 6	Phân phối chuẩn	110
6.1	Phân phối chuẩn	110
6.2	Các ứng dụng của phân phối chuẩn	113
6.2.1	Tìm giá trị dữ liệu được cho bởi xác suất cụ thể	115
6.2.2	Xác định tính chuẩn	116
6.3	Định lý giới hạn trung tâm	117
6.3.1	Phân phối của trung bình mẫu	118
6.3.2	Thừa số điều chỉnh hữu hạn	121
6.4	Phép tiệm cận chuẩn cho phân phối nhị thức	122
Chương 7	Khoảng tin cậy và cỡ mẫu	128
7.1	Khoảng tin cậy cho trung bình khi biết σ và cỡ mẫu	128
7.2	Khoảng tin cậy cho trung bình khi không biết σ	131
7.2.1	Độ lệch chuẩn tổng thể σ chưa biết, kích thước mẫu $n \geq 30$. .	131
7.2.2	Độ lệch chuẩn tổng thể σ chưa biết, kích thước mẫu $n < 30$, X tuân theo phân phối chuẩn	131
7.3	Khoảng tin cậy và cỡ mẫu cho tỷ lệ	134
7.4	Khoảng tin cậy cho phương sai và độ lệch chuẩn	136
Chương 8	Kiểm định giả thuyết thống kê	142
8.1	Thủ tục kiểm định giả thuyết – phương pháp truyền thống	142
8.2	Kiểm định Z cho giá trị trung bình	148
8.3	Kiểm định T cho giá trị trung bình	152
8.4	kiểm định Z cho tỉ lệ	153
8.5	Kiểm định χ^2 cho phương sai và độ lệch chuẩn	155
8.6	Các chủ đề khác liên quan đến kiểm định giả thuyết	157
8.6.1	Khoảng tin cậy và kiểm định giả thuyết	157

8.6.2	Sai lầm loại II và lực kiểm định	158
Chương 9	Tương quan và hồi quy	162
9.1	Đồ thị với các điểm chấm (<i>scatter plots</i>) và tương quan (<i>correlation</i>) .	163
9.2	Hồi quy (<i>regression</i>)	168
9.2.1	Đường hồi quy	168
9.2.2	Xác định phương trình đường hồi quy	169
9.3	Hệ số quyết định và sai số chuẩn tiên đoán	171
9.3.1	Các loại biến cho mô hình hồi quy	171
9.3.2	Hệ số quyết định (<i>coefficient of determination</i>)	173
9.3.3	Sai số tiêu chuẩn tiên đoán (<i>standard error of the estimate</i>) . .	173
9.3.4	Khoảng tiên đoán (<i>prediction interval</i>)	175
9.4	Hồi quy bội	176
9.4.1	Phương trình hồi quy bội	176
9.4.2	Kiểm định ý nghĩa của R	177
9.4.3	Điều chỉnh R^2	177
PHỤ LỤC		220
TÀI LIỆU THAM KHẢO		238

Chương 1

BẢN CHẤT CỦA XÁC SUẤT THỐNG KÊ

Thống kê là khoa học nghiên cứu của tập hợp nhiều lĩnh vực khác nhau, bao gồm thu thập, tổ chức, tổng hợp, phân tích và rút ra kết luận từ dữ liệu. Một vài lý do mà sinh viên học sinh cần học thống kê là:

- Giống như những người chuyên nghiệp, bạn phải có khả năng đọc và hiểu các nghiên cứu thống kê được thực hiện trong các lĩnh vực của bạn. Để có được sự hiểu biết này, bạn phải am hiểu về từ vựng, các kí hiệu, khái niệm, và thủ tục thống kê được sử dụng trong các nghiên cứu này.
- Bạn có thể được giao để tiến hành các nghiên cứu trong lĩnh vực của bạn, vì thủ tục thống kê là cơ bản để tiến hành một nghiên cứu. Để thực hiện việc này, bạn phải có khả năng thiết kế các thí nghiệm, thu thập, tổ chức, phân tích và tóm tắt dữ liệu và có thể đưa ra dự đoán đáng tin cậy hoặc dự báo để sử dụng trong tương lai. Bạn cũng phải có khả năng truyền đạt kết quả nghiên cứu bằng khả năng diễn đạt của bạn.
- Bạn cũng có thể sử dụng các kiến thức thu được từ nghiên cứu thống kê để trở thành người tiêu dùng và công dân tốt hơn. Ví dụ: bạn có thể đưa ra các quyết định thông minh về sản phẩm cần mua dựa trên nghiên cứu người tiêu dùng, về chi tiêu của chính phủ dựa trên nghiên cứu sử dụng,... Những lý do này có thể được coi là mục tiêu để nghiên cứu số liệu thống kê. Mục đích của chương này là giới thiệu các mục tiêu nghiên cứu thống kê bằng việc trả lời các câu hỏi như sau: Thống kê là gì? Dữ liệu là gì? Các mẫu được chọn như thế nào?

Sau khi học xong chương này, người học sẽ được cung cấp về những điều sau:

- Kiểm tra được kiến thức về các thuật ngữ thống kê.
- Phân biệt được giữa hai nhánh của thống kê: Thống kê mô tả và thống kê suy diễn.
- Xác định được các loại dữ liệu.
- Xác định mức đo lường cho mỗi biến.
- Xác định bốn kỹ thuật lấy mẫu căn bản.
- Giải thích sự khác nhau giữa nghiên cứu quan sát và nghiên cứu thực nghiệm.
- Giải thích cách mà thống kê có thể được sử dụng và sự lạm dụng của thống kê.
- Giải thích tầm quan trọng của máy tính và tính toán trong thống kê.

1.1 Thống kê mô tả và thống kê suy diễn

Để có được các kiến thức về những tình huống dường như ngẫu nhiên, thống kê thu thập thông tin cho các biến, mô tả tình hình. Một biến là một đặc điểm hay thuộc tính có thể cho giá trị khác nhau.

Dữ liệu là tập các giá trị (số đo hoặc quan sát) mà các biến có thể giả định. Biến có giá trị được xác định ngẫu nhiên được gọi là biến ngẫu nhiên. Tập các giá trị dữ liệu tạo thành một tập hợp giá trị dữ liệu. Mỗi giá trị dữ liệu thiết lập được gọi là một giá trị dữ liệu hoặc một số liệu.

Thống kê thường được chia ra hai loại là thống kê mô tả và thống kê suy diễn:

- + Thống kê mô tả được sử dụng để mô tả một nhóm phần tử được quan sát trong thực tế. Thống kê mô tả bao gồm thu thập, tổ chức, tổng hợp và trình bày dữ liệu.
- + Thông thường trong nghiên cứu, ta không chỉ giới hạn các kết luận vào một nhóm các phần tử mà ta đã quan sát (được gọi là mẫu). Điều mà ta muốn là tìm hiểu một sự liên hệ nào đó có thể áp dụng cho tất cả các phần tử mà ta chưa hay không thể quan sát được (được gọi là tổng thể). Quá trình tìm hiểu một mẫu rồi dựa vào các kết quả của lý thuyết xác suất để rút ra kết luận cho tổng thể được gọi là thống kê suy diễn. Như vậy mục đích chính của thống kê suy diễn là tổng hợp kết quả từ các mẫu cho tổng thể, thực hiện ước lượng và kiểm định giả thuyết, xác định các mối quan hệ giữa các biến và đưa ra dự đoán.

Tổng thể (còn được gọi là tập hợp chính, dân số) là tập hợp tất cả các phần tử do mục đích và phạm vi vấn đề cần nghiên cứu quy định. Mẫu là một nhóm các đối tượng được chọn trực tiếp từ tổng thể.

1.2 Các loại biến và các loại dữ liệu

Quan sát và đo lường các hiện tượng là điều căn bản cho tất cả các nghiên cứu khoa học. Các hiện tượng hoặc dấu hiệu mà ta quan sát có thể thay đổi từ phần tử này qua phần tử khác được gọi là **biến**. Ta cần phân biệt các loại biến: biến định tính (qualitative variables) và biến định lượng (quantitative variables). Các biến được đo lường và phân loại theo các thang đo, do đó ta cần phân biệt các loại thang đo: thang đo danh nghĩa (nominal-level), thang đo thứ hạng (ordinal-level), thang đo khoảng (interval-level) và thang đo tỉ lệ (ratio-level).

1.2.1 Biến định tính và biến định lượng

Biến định tính là những biến mà người ta gán các giá trị để phân biệt hay phân loại quan sát.

Chẳng hạn: giới tính (nam, nữ); tình trạng hôn nhân (độc thân, có gia đình, ly dị, góa chồng hay vợ), kết quả học tập (yếu, trung bình, khá, giỏi).

Biến định lượng là những biến mà các giá trị của nó được xác định bằng đo lường. Biến định lượng được chia thành hai loại: biến định lượng rời rạc và biến định lượng liên tục.

Chẳng hạn:

- a) Biến định lượng rời rạc như: số con trong một gia đình, số học sinh trong một lớp học,...
- b) Biến định lượng liên tục như: chiều cao; cân nặng; thời gian phản ứng; nhiệt độ,...

1.2.2 Các loại thang đo thường được sử dụng

Các biến được đo lường và phân loại theo 4 loại thang đo:

- i. **Thang danh nghĩa:** là thang đo gán các con số cho các quan sát để phân biệt và phân loại chúng, không có ý so sánh và các phép tính với chúng đều vô nghĩa. Ví dụ: giới tính nam được gán bởi số 0, nữ được gán bởi số 1.
- ii. **Thang thứ hạng:** là thang danh nghĩa mà giữa các dấu hiệu quan sát đã có quan hệ thứ bậc hơn kém. Ví dụ: thứ hạng học tập của sinh viên từ giỏi nhất đến kém nhất,... Trong thang đo này sự sai khác giữa các dấu hiệu quan sát không bắt buộc phải đều nhau.
- iii. **Thang đo khoảng:** là thang đo thứ hạng có khoảng cách đều nhau giữa các bậc. Gán các con số cho các quan sát phản ánh một chiều dài cố định giữa các đơn vị đo lường. Có thể đánh giá sự khác biệt giữa các dấu hiệu quan sát bằng loại thang đo này mặc dù điểm gốc ở đây chỉ là tương đối. Đây là thang đo có các khoảng cách đều nhau, việc cộng trừ các con số có ý nghĩa, có thể tính toán trên các con số của thang đo này. Để thu được thang đo khoảng có thể bắt đầu từ thang đo thứ hạng sau đó chuẩn hóa sao cho các khoảng cách đều nhau và việc tính toán các trị số đo trở nên có ý nghĩa. Chẳng hạn, để đặc trưng lứa tuổi có thể dùng thang đo khoảng: trẻ (dưới 35 tuổi) [30], trung niên (từ 36 tuổi đến 60 tuổi) [50], già (từ 60 tuổi trở lên) [70]. Các thang đo danh nghĩa, thứ hạng và thang đo khoảng dùng để đặc trưng các giá trị của dấu hiệu định tính.
- iv. **Thang tỉ lệ:** là thang đo khoảng cách với một điểm 0 tuyệt đối (điểm gốc) để có thể so sánh được tỉ lệ giữa các trị số đo. Với thang đo này ta có thể đo lường các dấu hiệu quan sát và thực hiện được tất cả phép tính với trị số đo. Thang đo tỉ lệ dùng để đặc trưng các giá trị của dấu hiệu định lượng. Theo tuần tự thang đo sau có chất lượng cao hơn thang đo trước, mỗi thang đo cấp cao hơn có thể chuyển xuống thang đo cấp thấp hơn.

Một số ví dụ về các thang đo:

Thang danh nghĩa	Thang thứ hạng	Thang đo khoảng	Thang tỉ lệ
Mã bưu chính	Điểm số (A,B,C,D,F)	Điểm SAT	Chiều cao
Giới tính (nam, nữ)	Đánh giá (vị trí nhất, nhì, ...)	Chỉ số IQ	Cân nặng
Màu mắt (nâu, xanh dương, ...)	Thang điểm đánh giá (yếu, khá, ...)	Nhiệt độ	Thời gian
Đảng phái chính trị	Xếp hạng các cầu thủ		Lương
Chuyên ngành (toán, tin học, ...)	quần vợt		Tuổi
Quốc tịch			
Tôn giáo			

1.3 Thu thập dữ liệu và phương pháp lấy mẫu

Dữ liệu có thể được thu thập bằng nhiều cách khác nhau. Một trong những phương pháp phổ biến nhất là thông qua việc sử dụng các cuộc điều tra. Khảo sát có thể được thực hiện bằng cách sử dụng nhiều phương pháp. Ba trong số các phương pháp phổ biến nhất là khảo sát qua điện thoại, bản câu hỏi gửi qua đường bưu điện và cuộc phỏng vấn cá nhân.

Các cuộc điều tra bằng điện thoại có lợi thế hơn các cuộc điều tra phỏng vấn cá nhân do họ ít tốn kém hơn. Ngoài ra, người ta có thể thẳng thắn hơn trong quan điểm của họ vì không có liên hệ đối mặt. Một trở ngại lớn nhất đối với cuộc khảo sát qua điện thoại là một số người dân sẽ không có điện thoại hoặc sẽ không trả lời khi có cuộc gọi; do đó, không phải tất cả mọi người đều có cơ hội được điều tra. Ngoài ra, hiện nay nhiều người dùng số điện thoại không có trong danh sách công bố, vì vậy họ không thể được khảo sát. Cuối cùng, ngay cả giọng nói của người phỏng vấn cũng có thể ảnh hưởng đến phản ứng của người được phỏng vấn.

Các cuộc điều tra gửi qua đường bưu điện có thể được sử dụng để bao phủ một khu vực địa lý rộng lớn hơn so với các cuộc điều tra qua điện thoại hoặc các cuộc phỏng vấn cá nhân vì các cuộc điều tra bằng phiếu gửi ít tốn kém hơn để tiến hành. Ngoài ra, người trả lời có thể vẫn vô danh nếu họ muốn. Nhược điểm của các cuộc điều tra bằng bảng câu hỏi gửi thư bao gồm số lượng phản hồi thấp và câu trả lời không thích hợp. Một nhược điểm nữa là một số người có thể gặp khó khăn khi đọc hoặc hiểu các câu hỏi.

Các cuộc điều tra phỏng vấn cá nhân có lợi thế là có được những câu trả lời sâu về các câu hỏi từ người được phỏng vấn. Một bất lợi là người phỏng vấn phải được đào tạo trong việc đặt câu hỏi và ghi lại những phản hồi, làm cho cuộc điều tra phỏng vấn cá nhân tốn kém hơn so với các phương pháp khảo sát khác. Một bất lợi khác là người phỏng vấn có thể có thành kiến trong việc lựa chọn người trả lời của mình.

Dữ liệu cũng có thể được thu thập theo những cách khác, chẳng hạn như khảo sát hồ sơ hoặc quan sát trực tiếp các tình huống.

Các nhà nghiên cứu sử dụng các mẫu để thu thập dữ liệu và thông tin về một biến cụ thể từ một quần thể lớn. Sử dụng mẫu tiết kiệm về mặt thời gian, tiền bạc và trong một số trường hợp cho phép nhà nghiên cứu có được thông tin chi tiết hơn về một chủ đề cụ thể. Tuy nhiên các mẫu không thể được lựa chọn theo cách ngẫu

nhiên được bởi vì thông tin thu được có thể là chệch lệch. Để có được các mẫu không chệch lệch - nghĩa là mỗi đối tượng trong quần thể có cơ hội được lựa chọn như nhau - các nhà thống kê sử dụng bốn phương pháp lấy mẫu căn bản: mẫu ngẫu nhiên, có hệ thống, phân tầng và lấy mẫu cụm.

1.3.1 Mẫu ngẫu nhiên (*Random Sampling*)

Các mẫu ngẫu nhiên được lựa chọn bằng cách sử dụng các phương pháp ngẫu nhiên hoặc các số ngẫu nhiên. Một trong những phương pháp như vậy là gán mỗi số cho một đối tượng trong tổng thể. Sau đó đặt các thẻ được đánh số vào một cái bát, trộn chúng cẩn thận và chọn nhiều loại thẻ nếu cần. Các đối tượng có số được chọn tạo thành mẫu. Vì khó trộn các loại thẻ này kỹ lưỡng nên vẫn có thể tạo ra một mẫu chệch. Vì lý do này, các nhà thống kê sử dụng phương pháp khác để thu thập số liệu bằng cách họ tạo ra các số ngẫu nhiên từ máy tính.

1.3.2 Mẫu hệ thống (*Systematic Sampling*)

Là loại mẫu đã được đơn giản hóa trong cách chọn, trong đó chỉ có phần tử đầu tiên được chọn ngẫu nhiên, sau đó dựa vào danh sách đã được đánh số của tổng thể để chọn ra các phần tử tiếp theo vào mẫu theo một thủ tục nào đó. Chẳng hạn, trên một danh sách N phần tử cần chọn ra một mẫu kích thước n thì ta chia danh sách đó ra n phần bằng nhau, ở phần thứ nhất gồm N/n phần tử, chọn ngẫu nhiên ra một phần tử, sau đó theo danh sách cứ cách N/n phần tử ta lấy ra một phần tử vào mẫu cho đến khi có đủ n phần tử.

1.3.3 Mẫu phân tầng (*Stratified Sampling*)

Để thu được một mẫu phân tầng ta phân chia tổng thể thành các nhóm (gọi là tầng) theo một số đặc điểm quan trọng cho nghiên cứu, sau đó chọn ngẫu nhiên các phần tử đại diện cho từng nhóm.

1.3.4 Mẫu chùm (*Cluster Sampling*)

Trong một số trường hợp để tiện cho việc nghiên cứu người ta muốn quy diện nghiên cứu gọn về một khu vực nhất định chứ không để cho các phần tử của mẫu phân tán quá rộng, lúc đó mẫu được chọn theo chùm.

Để thực hiện theo phương pháp này, trước tiên tổng thể điều tra được phân chia thành nhiều chùm theo nguyên tắc:

- mỗi phần tử của tổng thể chỉ được phân vào một chùm.
- mỗi chùm cố gắng chứa nhiều phần tử khác nhau về dấu hiệu nghiên cứu sao cho nó có độ phân tán cao như tổng thể.
- phân chia sao cho các chùm tương đối đồng đều nhau về quy mô.
Tiếp đó các chùm được chọn một cách ngẫu nhiên và tất cả các phần tử của chùm đó đều được chọn vào mẫu.

Tóm tắt phương pháp lấy mẫu:

Mẫu ngẫu nhiên	Đối tượng được lựa chọn theo số ngẫu nhiên.
Mẫu hệ thống	Các đối tượng được lựa chọn cách nhau k lần sau khi đối tượng đầu tiên được chọn ngẫu nhiên từ 1 đến k .
Mẫu phân tầng	Các đối tượng được lựa chọn bằng cách phân chia tổng thể thành các nhóm (tầng) và các đối tượng được chọn ngẫu nhiên trong các nhóm.
Mẫu chùm	Các đối tượng được lựa chọn bằng cách sử dụng một nhóm nguyên vẹn đại diện cho quần thể.

1.4 Nghiên cứu quan sát và thực nghiệm

1.4.1 Nghiên cứu quan sát

Trong một nghiên cứu quan sát, nhà nghiên cứu chỉ quan sát thấy những gì đang xảy ra hoặc những gì đã xảy ra trong quá khứ và cố gắng rút ra kết luận dựa trên những quan sát này.

1.4.2 Nghiên cứu thực nghiệm

Thực nghiệm là một phương pháp thu thập thông tin được thực hiện bởi những quan sát trong điều kiện gây biến đổi đối tượng khảo sát và môi trường xung quanh đối tượng khảo sát một cách có chủ định. Phương pháp thực nghiệm được áp dụng phổ biến không chỉ trong nghiên cứu tự nhiên, kỹ thuật, y học mà cả trong xã hội và các lĩnh vực khác.

Thực nghiệm cho phép tác động lên đối tượng nghiên cứu một cách chủ động, can thiệp có ý thức vào quá trình diễn biến tự nhiên, để hướng quá trình diễn ra theo mong muốn của nhà nghiên cứu.

Các nghiên cứu thống kê thường bao gồm một hoặc nhiều biến độc lập và một biến phụ thuộc. Các biến được sử dụng để mô tả hoặc đo lường vấn đề nghiên cứu gọi là biến phụ thuộc (dependent variable). Các biến được sử dụng để mô tả hoặc đo lường các yếu tố (tác nhân) được giả định là gây ra hoặc ít nhất là làm ảnh hưởng đến vấn đề nghiên cứu được gọi là biến độc lập (independent variable).

Ví dụ: trong một nghiên cứu về mối liên quan giữa hút thuốc lá và ung thư phổi thì biến “có bị ung thư phổi hay không” (nhận các giá trị có hoặc không) sẽ là biến phụ thuộc, còn biến “hút thuốc” (biến thiên từ không hút thuốc đến hút trên 3 bao một ngày) là biến độc lập.

1.5 Lợi ích và lạm dụng của thống kê

Thống kê có thể được sử dụng để mô tả dữ liệu, so sánh hai hoặc nhiều tập dữ liệu, xác định mối liên hệ giữa các biến, kiểm định giả thuyết và đưa ra các ước tính

về đặc điểm của tổng thể. Tuy nhiên, có một khía cạnh khác của số liệu thống kê đó là việc sử dụng sai kỹ thuật thống kê để bán sản phẩm không hoạt động đúng cách để thử chứng minh điều gì đó thực sự là không đúng sự thật hoặc để thu hút sự chú ý của chúng ta bằng cách sử dụng thống kê để gây ra nỗi sợ hãi, sốc và xúc phạm.

Sau đây là một số cách mà thống kê có thể bị trình bày sai:

- Sử dụng mẫu nghi ngờ: Đôi khi các nhà nghiên cứu sử dụng các mẫu rất nhỏ để có được thông tin. Không chỉ quan trọng là phải có kích thước mẫu đủ lớn mà còn là cần thiết để xem các đối tượng trong mẫu đã được lựa chọn như thế nào. Một số nhà nghiên cứu dùng mẫu thuận tiện. Chẳng hạn, các nghiên cứu giáo dục đôi khi sử dụng toàn bộ học sinh trong một lớp học vì thuận tiện để đánh giá cho toàn bộ tổng thể.
- Trung bình không được rõ ràng.
- Một biến dạng của thống kê có thể xảy ra khi các giá trị khác nhau được biểu diễn cho cùng một dữ liệu.
- Thống kê bị tách rời, tức là thống kê không có sự so sánh được thực hiện.
- Nhiều kết nối hàm ý giữa các biến mà có thể không thực sự tồn tại. Ví dụ, hãy xem xét tuyên bố sau: "Ăn cá có thể giúp giảm cholesterol của bạn". Lưu ý các từ "có thể giúp". Không có gì đảm bảo rằng ăn cá chắc chắn sẽ giúp bạn giảm cholesterol.
- Sử dụng biểu đồ sai lệch sẽ gây cho người đọc rút ra những kết luận sai.
- Sử dụng các câu hỏi khảo sát bị lỗi.

1.6 Máy tính và tính toán

Trong phần này, tôi giới thiệu một vài phần mềm được sử dụng trong xác suất thống kê: SPSS, MINITAB, Microsoft Excel, Máy tính bỏ túi.

Phần bài tập chương 1

1. Nêu tên và định nghĩa hai lĩnh vực của thống kê.
2. Giải thích sự khác nhau giữa mẫu và tổng thể.
3. Tại sao mẫu lại được sử dụng trong thống kê?
4. Trong mỗi báo cáo sau, thống kê mô tả hay thống kê suy diễn được sử dụng?
 - a. Trong năm 2010, 148 triệu người Mỹ sẽ tham gia HMO (Nguồn: USA TODAY).
 - b. Chín trong số mười người tử vong trong công việc là nam giới (Nguồn: USA TODAY Weekend).
 - c. Chi phí cho ngành công nghiệp cáp là 5,66 tỷ đô la vào năm 1996 (Nguồn: USA TODAY).
 - d. Thu nhập trung bình của hộ gia đình cho người từ 25-34 tuổi là 35.888 USD (Nguồn: USA TODAY).
 - e. Liệu pháp dị ứng làm cho ong bỏ đi (Nguồn: Phòng ngừa).
 - f. Uống cà phê không có caffein có thể làm tăng mức cholesterol lên 7% (Nguồn: Hiệp hội Tim Mạch Hoa Kỳ).
 - g. Chi phí y tế trung bình hàng năm cho mỗi người là 1052 đô la (Nguồn: The Greensburg Tribune Review).
 - h. Các chuyên gia nói rằng tỷ lệ thể chấp có thể sớm hạ xuống đến mức thấp nhất (Nguồn: USA TODAY).
5. Phân loại từng thang đo: thang đo danh nghĩa, thang thứ hạng, thang đo khoảng, thang tỉ lệ cho các dữ kiện sau:
 - a. Số trang trong cuốn danh bạ điện thoại của thành phố Cleveland.
 - b. Xếp hạng của cầu thủ quần vợt.
 - c. Trọng lượng của các máy điều hòa.
 - d. Nhiệt độ bên trong 10 tủ lạnh.
 - e. Tiền lương của năm giám đốc điều hành hàng đầu tại Hoa Kỳ.
 - f. Xếp hạng của tám vở kịch địa phương (yếu, trung bình, tốt, xuất sắc).
 - g. Thời gian cần thiết cho thợ cơ khí để điều chỉnh máy.
 - h. Tuổi của học sinh trong lớp.
 - i. Tình trạng hôn nhân của bệnh nhân tại văn phòng bác sĩ.
 - j. Mã lực của động cơ máy kéo.
6. Phân loại các biến sau theo biến định tính hay định lượng.
 - a. Số lượng xe đạp được bán trong 1 năm bởi một cửa hàng bán đồ thể thao lớn.
 - b. Màu sắc của mũ bóng chày trong một cửa hàng.
 - c. Thời gian cần thiết để cắt một bãi cỏ.

- d. Dung tích (feet khối) của sáu chiếc xe tải.
- e. Phân loại trẻ em ở trung tâm chăm sóc ban ngày (trẻ sơ sinh, trẻ mới biết đi, mẫu giáo).
- f. Trọng lượng cá bắt được ở Hồ George.
- g. Tình trạng hôn nhân của các giảng viên trong một trường đại học lớn.

7. Phân loại biến rời rạc và biến liên tục:

- a. Số lượng bánh rán được bán hàng ngày bởi Donut Heaven.
- b. Nhiệt độ nước của sáu hồ bơi ở Pittsburgh vào một ngày nhất định.
- c. Trọng lượng của các con mèo trong nơi trú ẩn vật nuôi.
- d. Tuổi thọ (tính bằng giờ) 12 pin của đèn pin.
- e. Số lượng bánh mì kẹp bơ được bán mỗi ngày bởi một cửa hàng Hamburger Stand trong khuôn viên một trường đại học.
- f. Số lượng DVD được thuê mỗi ngày bởi một cửa hàng video.
- g. Dung tích (gallon) của sáu hồ chứa ở Hạt Jefferson.

8. Nêu tên và định nghĩa bốn phương pháp lấy mẫu cơ bản.

9. Phân loại từng mẫu sau theo: mẫu ngẫu nhiên, mẫu hệ thống, mẫu phân tầng, mẫu chùm.

- a. Trong một trường học lớn thuộc quận, tất cả giáo viên từ hai tòa nhà được phỏng vấn để xác định xem họ tin rằng các sinh viên có ít bài tập ở nhà phải làm bấy giờ hơn những năm trước.
- b. Mỗi thứ bảy, các khách hàng vào khu mua sắm được yêu cầu chọn cửa hàng yêu thích của mình.
- c. Giám sát viên điều dưỡng được chọn tùy ý một số để xác định mức lương hàng năm.
- d. Mỗi bánh hamburger thứ 100 được sản xuất đều được kiểm tra để xác định hàm lượng chất béo của nó.
- e. Người vận chuyển thư của một thành phố lớn được chia thành bốn nhóm theo giới tính (nam hay nữ) và tùy theo họ đi bộ hay đi xe trên các tuyến đường của họ. Sau đó 10 người được lựa chọn từ mỗi nhóm và phỏng vấn để xác định xem họ đã bị chó cắn vào năm ngoái hay không.

10. Với mỗi câu dưới đây, hãy xác định tổng thể và nêu rõ cách lấy mẫu.

- a. Chi phí trung bình của một bữa ăn hàng không là 4,55 đô la (Nguồn: Everything Has Its Price, Richard E. Donley, Simon and Schuster).
- b. Hơn 1 trong 4 trẻ em ở Hoa Kỳ có mức cholesterol ở mức 180 miligam hoặc cao hơn (Nguồn: Quỹ y tế Hoa Kỳ).
- c. Mỗi 10 phút, 2 người chết vì tai nạn xe hơi và 17 người bị thương (Nguồn: ước tính của Hội đồng An toàn Quốc gia).

- d. Khi người già với cao huyết áp từ nhẹ đến trung bình được truyền muối khoáng trong vòng 6 tháng, chỉ số huyết áp trung bình giảm 8 điểm tâm thu và tâm trương 3 điểm (Nguồn: Phòng bệnh).
- e. Số tiền trung bình dành cho mỗi món quà cho Mẹ trong Ngày của Mẹ là 25,95 đô la (Nguồn: Tổ chức Gallup).

11. Xác định các nghiên cứu sau là nghiên cứu quan sát hay nghiên cứu thực nghiệm?

- a. Các đối tượng được phân ngẫu nhiên vào hai nhóm, và một nhóm được cho một loại thảo mộc và một nhóm khác là giả dược. Sau 6 tháng, số người mắc bệnh đường hô hấp trên mỗi nhóm đã được so sánh.
- b. Một nhà nghiên cứu đứng ở một ngã tư đông đúc để xem liệu màu của ô tô có liên quan đến việc người lái vượt đèn đỏ.
- c. Một nhà nghiên cứu nhận thấy rằng những người gây hấn hơn sẽ có mức cholesterol toàn phần cao hơn những người ít gây hấn.
- d. Các đối tượng được phân chia ngẫu nhiên thành bốn nhóm. Mỗi nhóm được xếp vào một trong bốn chế độ ăn đặc biệt: chế độ ăn ít chất béo, chế độ ăn nhiều cá, sự kết hợp giữa chế độ ăn ít chất béo và chế độ ăn nhiều cá, chế độ ăn chuẩn. Sau 6 tháng, huyết áp của các nhóm được so sánh để xem chế độ ăn uống có bất kỳ ảnh hưởng nào đến huyết áp không.

12. Xác định các biến độc lập và biến phụ thuộc trong mỗi nghiên cứu trong ví dụ 11

13. **Các vi khuẩn có lợi** Theo một nghiên cứu thí điểm của 20 người được tiến hành tại Đại học Minnesota, hàng ngày cho sử dụng thuốc theo liều lượng của một hợp chất được gọi là arabinogalactan trong thời gian 6 tháng đã làm tăng đáng kể các loại vi khuẩn lactobacillus có lợi. Tại sao không thể kết luận rằng hợp chất này có lợi cho đa số mọi người?

Chương 2

PHÂN BỐ TẦN SỐ VÀ ĐỒ THỊ

Khi tiến hành nghiên cứu thống kê, nhà nghiên cứu phải thu thập dữ liệu cho một biến cụ thể đang được nghiên cứu. Ví dụ, nếu một nhà nghiên cứu muốn nghiên cứu số người bị rắn cắn ở một khu vực địa lý cụ thể trong vài năm gần đây, họ phải thu thập dữ liệu từ các bác sĩ, bệnh viện hoặc các sở y tế khác nhau.

Để mô tả tình huống, rút ra kết luận, hoặc suy luận về sự kiện, nhà nghiên cứu phải tổ chức dữ liệu một cách có ý nghĩa. Cách thức thuận tiện nhất để tổ chức dữ liệu là xây dựng một phân bố tần số. Sau khi tổ chức dữ liệu, nhà nghiên cứu phải trình bày cho người đọc có thể rút ra một số thông tin từ thị giác từ những sự kiện có vẻ như phức tạp, hỗn độn. Phương pháp hữu ích nhất để trình bày dữ liệu là xây dựng biểu đồ và đồ thị thống kê. Có rất nhiều loại biểu đồ và đồ thị khác nhau và mỗi loại có một mục đích cụ thể.

Chương này giải thích làm thế nào để tổ chức dữ liệu bằng cách xây dựng phân phối tần số và làm thế nào để trình bày dữ liệu bằng cách xây dựng biểu đồ và đồ thị. Các biểu đồ và biểu đồ được minh họa ở đây là nhật đồ, đa giác tần số, biểu đồ hình cung, biểu đồ hình tròn, biểu đồ Pareto và biểu đồ chuỗi thời gian. Một biểu đồ kết hợp các đặc điểm của một phân bố tần số và nhật đồ được gọi là biểu đồ thân và lá.

2.1 Tổ chức dữ liệu

Giả sử một nhà nghiên cứu muốn nghiên cứu về lứa tuổi của 50 người giàu nhất trên thế giới. Đầu tiên các nhà nghiên cứu sẽ phải lấy dữ liệu về tuổi của các người dân. Trong trường hợp này, những lứa tuổi này được liệt kê trong tạp chí Forbes. Khi dữ liệu ở dạng ban đầu, chúng được gọi là dữ liệu thô và được liệt kê dưới đây:

49	57	38	73	81
74	59	76	65	69
54	56	69	68	78
65	85	49	69	61
48	81	68	37	43
78	82	43	64	67
52	57	81	77	79
85	40	85	59	80
60	71	57	61	69
61	83	90	87	74

Vì khi xem dữ liệu thô ta thu được ít thông tin từ nó nên nhà nghiên cứu thiết lập một phân bố tần số cho dữ liệu. Phân bố tần số là tổ chức dữ liệu thô ở dạng bảng bao gồm các lớp và tần số tương ứng. Số lần xảy ra cho mỗi lớp được gọi là tần

số (frequency) của lớp đó. Nếu thực hiện công việc này bằng tay, ta ghi dấu (tally) mỗi lần xảy ra cho mỗi loại, bằng cách như vậy ta thiết lập một phân bố tần số cho mẫu của ta. Tuy nhiên nếu chỉ có tần số mà thôi thì chưa nói lên được điều gì nếu ta không cải biến nó ra tỉ lệ phần trăm. Tần suất của một lớp là tỉ lệ phần trăm của lớp đó.

Class limits	Tally	Frequency
35–41	///	3
42–48	///	3
49–55	////	4
56–62	 	10
63–69	 	10
70–76		5
77–83	 	10
84–90		5
		Total 50

Từ bảng phân bố tần số, ta có thể nói rằng đa số những người giàu có trong nghiên cứu này trên 55 tuổi.

a. Đối với biến định tính

Phân bố tần số phân loại (*categorical frequency distribution*) được sử dụng cho biến định tính.

Quy trình:

- Bước 1: Lập bảng như sau:

Lớp (Class)	Ghi dấu (Tally)	Tần số (Frequency)	Phần trăm (Percent)

- Bước 2: Ghi dấu dữ liệu.
- Bước 3: Tìm tần số và tỉ lệ phần trăm.

Ví dụ 2.1.1. Hai mươi lăm cảnh sát viên được xét nghiệm máu để xác định nhóm máu của họ. Tập dữ liệu được cho như sau:

A	B	B	AB	O
O	O	B	AB	B
B	B	O	A	O
A	O	O	O	AB
AB	A	O	B	A

Hãy xây dựng một phân bố tần số cho dữ liệu trên.

Lời giải:

A Class	B Tally	C Frequency	D Percent
A		5	20
B		7	28
O		9	36
AB		4	16
		Total 25	100

Như vậy, đối với mẫu trên thì nhiều người có loại máu O hơn bất kỳ loại nào khác.

b. Đối với biến định lượng

Với biến định lượng có hai loại phân bố tần số sau:

i) Phân bố tần số ghép lớp (*grouped frequency distribution*):

Phân bố tần số ghép lớp được sử dụng khi phạm vi dữ liệu (*range*) lớn, mỗi lớp là một đoạn dữ liệu có độ rộng (*width*) lớn hơn 1 đơn vị.

Quy trình phân bố tần số ghép lớp:

- Bước 1: Xác định các lớp.

- + Tìm giá trị lớn nhất, kí hiệu H (*Highest*) và giá trị nhỏ nhất, kí hiệu L (*Lowest*).
- + Tìm phạm vi dữ liệu (*range*): $R = H - L$.
- + Chọn số lớp theo mong muốn, kí hiệu NC .
- + Tìm độ rộng (*width*): $W = \frac{R}{NC}$. Với lưu ý: Nếu $W \notin \mathbb{Z}$ thì ta làm tròn lên và nếu $W \in \mathbb{Z}$ thì chọn độ rộng là $W + 1$.
- + Chọn điểm bắt đầu (giá trị nhỏ hơn hoặc bằng với giá trị nhỏ nhất của dữ liệu) để làm giới hạn dưới (*lower limit*), kí hiệu LL , của lớp đầu tiên, cộng thêm độ rộng để được các giới hạn dưới tiếp theo, tức là: $LL_{i+1} = LL_i + W$, trong đó i là chỉ số lớp thứ i .
- + Tìm các giới hạn trên (*upper limit*) của các lớp, kí hiệu là $UL_i = LL_i + W - 1$.
- + Tìm các lớp cận biên (*class boundaries*). Cận biên dưới (*lower class boundary*) của lớp thứ i , kí hiệu là LCB , cận biên trên (*upper class boundary*) của lớp thứ i , kí hiệu là UCB . Công thức xác định các lớp cận biên:

$$LCB_i = LL_i - \frac{1}{2} (\text{đơn vị đo lường dữ liệu}),$$

$$UCB_i = UL_i + \frac{1}{2} (\text{đơn vị đo lường dữ liệu}).$$

- Bước 2: Ghi dấu dữ liệu.

- Bước 3: Tìm các tần số và phần trăm.

Khi lập bảng phân bố tần số ghép lớp ta cần lưu ý:

- Có bao nhiêu lớp cần sử dụng trong phân bố? Số lớp nên có từ 5 đến 20 mặc dù không có một quy tắc cố định về số lớp trong một phân bố tần số, điều quan trọng nhất là phải có đủ các lớp để trình bày rõ ràng về các dữ liệu thu thập được. Vì nếu số lớp quá nhiều thì lợi ích của việc phân bố tần số ghép lớp không được bao nhiêu so với các dữ liệu thô. Còn nếu số lớp quá ít, nhiều điểm số được gộp vào một lớp, như vậy sẽ mất nhiều thông tin.
- Trong thống kê cơ bản thì các lớp có độ rộng là như nhau. Nhưng trong thống kê chuyên ngành thì độ rộng của các lớp có thể không bằng nhau tùy theo mục đích của nghiên cứu.

Ví dụ 2.1.2. Cho các số liệu thống kê về nhiệt độ cao kỷ lục (tính bằng độ F) của 50 tiểu bang. Hãy lập bảng phân bố tần số ghép lớp với 7 lớp.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Giải:

- Bước 1: Xác định các lớp:
 - + $H = 134, L = 100$.
 - + $R = H - L = 34$.
 - + $NC = 7$.
 - + $W = \frac{34}{7} = 4.9$ do đó ta chọn $W = 5$.
 - + Xác định các lớp giới hạn:

$$LL_1 = L = 100; LL_2 = 105; LL_3 = 110; LL_4 = 115; LL_5 = 120; LL_6 = 125; LL_7 = 130;$$

$$UL_1 = 104; UL_2 = 109; UL_3 = 114; UL_4 = 119; UL_5 = 124; UL_6 = 129; UL_7 = 134.$$
 - + Xác định các lớp cận biên:

$$LCB_1 = 99.5; LCB_2 = 104.5; \dots$$

$$UCB_1 = 104.5; UCB_2 = 109.5; \dots$$
- Bước 2: Ghi dấu dữ liệu.
- Bước 3: Tìm các tần số của mỗi lớp và ta được thống kê như sau:

Class limits	Class boundaries	Tally	Frequency
100–104	99.5–104.5	//	2
105–109	104.5–109.5	///	3
110–114	109.5–114.5	///	18
115–119	114.5–119.5	///	13
120–124	119.5–124.5	///	7
125–129	124.5–129.5	/	1
130–134	129.5–134.5	/	1
			$n = \Sigma f = 50$