THỐNG KÊ MÁY TÍNH

(Computational Statistics)

Trường Đại học Nha Trang
Khoa Công nghệ thông tin
Bộ môn Hệ thống thông tin
Giảng viên: TS.Nguyễn Khắc Cường

CHƯƠNG 2

THỐNG KÊ HỌC

2.1.Thống kê học

- Thống kê?
  - Thu thập dữ liệu
  - Mô tả dữ liệu

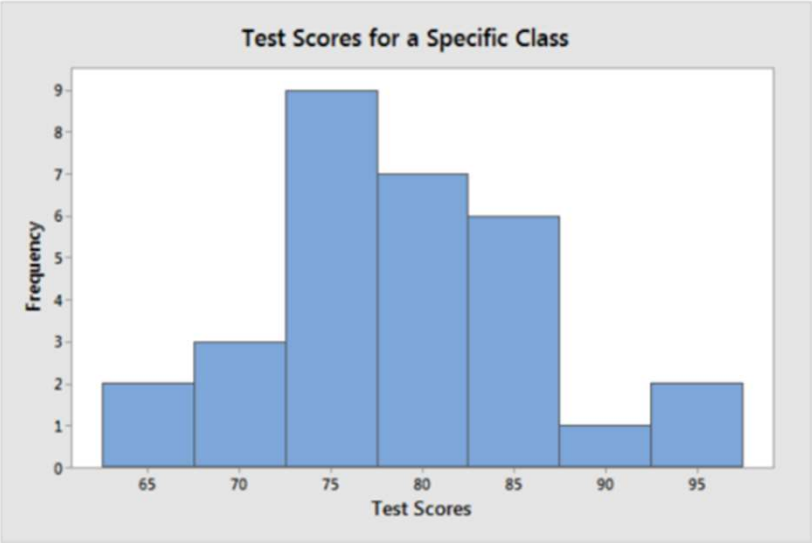
  → Tác dụng:
    - Rút ra qui luật tồn tại/chi phối trong dữ liệu/hiện tượng quan sát
    - Làm cơ sở để đưa ra quyết định nào đó
- Cơ sở lý thuyết của thống kê
  - Lý thuyết xác suất
  - Lý thuyết thống kê toán
- Ứng dụng
  - Thống kê trong dân số, kinh doanh, xã hội, y học, giáo dục, … → Hiểu? Dự đoán? Đưa ra quyết định?

2.1.Thống kê học

- Phân loại
  - Thống kê mô tả (Descriptive Statistics)
    - Là các phương pháp tóm tắt, mô tả dữ liệu
    - Gồm các phương pháp:
      - Thu thập dữ liệu; Phân tích dữ liệu
      - Tóm tắt, trình bày dữ liệu / kết quả phân tích
  - Thống kê suy luận (Inferential Statistics)
    - Là các phương pháp mô hình hóa dữ liệu để giải thích sự thay đổi ngẫu nhiên và không chắc chắn của dữ liệu → nhằm tạo ra các suy diễn đối với dữ liệu
    - Gồm các phương pháp:
      - Ước lượng thống kê; Kiểm định giả thuyết thống kê; Phân tích phương sai; Hồi qui; Dự báo thống kê; . . .

2.1.Thống kê học

- Ví dụ:
  - Thống kê mô tả (Descriptive Statistics)
    - Khảo sát và mô tả thông tin liên quan đến điểm thi của một lớp học



Statistic	Class value
Mean	79.18

Standard deviation	7.76
Proportion >= 70	86.7%

  - Thống kê suy luận (Inferential Statistics)
    - Dựa vào kết quả thi của vài lớp → Rút ra các kết luận, dự đoán về tình hình / xu hướng thay đổi điểm thi của toàn trường

2.2.Các khái niệm cơ bản

- Các bước cơ bản của nghiên cứu thống kê
  - Xác định vấn đề, mục tiêu, nội dung, đối tượng nghiên cứu
  - Xây dựng hệ thống các khái niệm, chỉ tiêu thông kê
  - Thu thập dữ liệu
  - Xử lý số liệu (kiểm tra, biến đổi, sắp xếp, . . .)
  - Phân tích
  - Trình bày / Giải thích các kết quả nghiên cứu

2.2.Các khái niệm cơ bản

- Population (quần thể / tổng thể)
  - Tập hợp các phần tử thuộc dữ liệu / đối tượng cần
    - Quan sát
    - Thu thập
    - Phân tích
  - Ví dụ:
    - Cần nghiên cứu mức độ sử dụng tiếng Anh trong học tập, nghiên cứu của sinh viên ĐH Nha Trang
    - Vậy:
      - Quần thể = toàn bộ sinh viên ĐH Nha Trang

2.2.Các khái niệm cơ bản

- Sample (mẫu)
  - Là một tập con gồm các phần tử của một quần thể
  - Các phần tử được chọn theo một phương pháp nào đó
  - Các phép xử lý / phân tích được thực hiện đối với samples
  - Từ các hiểu biết đối với các samples → suy diễn tổng quát lên để hiểu về population
- Observation (quan sát)
  - Một đơn vị của sample là một quan sát
  - Ví dụ:

	Weight (pounds)	Length (inches)	Region
Observation #1	290	30	East
	296	35	East
	299	34	East
	300	34	East
	305	38	East
	307	40	North
	311	46	North
	315	45	North
	325	49	North
	339	48	North
	340	55	South
	355	58	South
	357	55	West
	359	57	West
	361	59	West

2.2.Các khái niệm cơ bản

- Characteristic (đặc điểm thống kê)
  - Là đặc điểm liên quan đến nội dung nghiên cứu
  - Tương ứng với thuộc tính / tập thuộc tính trong database
  - Phân loại:
    - Qualitative characteristic (đặc điểm định tính)
      - Không thể biểu diễn đặc điểm bằng giá trị số
      - VD: tôn giáo, thói quen, . . .
    - Quantitative characteristic (đặc điểm định tính)
      - Đặc điểm biểu diễn được bằng giá trị số
        - Liên tục
        - Rời rạc
      - VD: rời rạc (mức lương, tuổi , . . .)
        liên tục (thời gian, nhiệt độ, …)

2.2.Các khái niệm cơ bản

- Data type (kiểu dữ liệu)
  - Nominal (định danh)
    - Các nhãn (label), mục (category) dùng mô tả, phân loại đối tượng
    - VD: mã hàng hóa, tên hàng hóa, . . .
  - Binary (nhị phân)
    - Là dữ liệu kiểu nominal nhưng chỉ mang một trong hai giá trị
    - VD: đúng/sai; nam/nữ; . . .
  - Ordinary (thứ tự)
    - Các phần tử của dữ liệu được sắp xếp có thứ tự
    - VD: xếp loại (kém, trung bình, khá, giỏi)
  - Interger (số nguyên)
    - Các phần tử của dữ liệu có kiểu số nguyên (mệnh giá tiền, năm,…)
    - Các toán tử tác động vào các phần tử để tạo ra phần tử mới

2.2.Các khái niệm cơ bản

- Data type (kiểu dữ liệu)
  - Interval (khoảng)
    - Các phần tử có giá trị cách đều nhau, thuộc một khoảng giá trị
    - Thường dùng làm thang đo
    - VD: vận tốc đo theo km/h, nhiệt độ đo theo độ C, . . .
  - Ratio-scaled (Khoảng tỉ lệ)
    - Tương tự như interval
    - Giá trị của các phần tử có mối liên hệ là bội số / ước số của nhau
    - VD: tuổi (20-25), (25-30), (lớn hơn 30)

2.3.Các kỹ thuật lấy mẫu

- Giới thiệu:
  - Thực tế:
    - Thường không thể thu thập được toàn bộ các phần tử / quan sát của population
    - Do đó, cần chọn các phần tử / quan sát làm đại diện

      → gọi là lấy mẫu
  - Yêu cầu:
    - Các mẫu được chọn phải là mẫu đại diện cho quần thể, phản ánh càng đúng càng tốt tính chất vốn có của quần thể

      → do đó, các kỹ thuật lấy mẫu đúng đắn là rất quan trọng
  - Ý nghĩa:
    - Lấy mẫu càng tốt / chất lượng cao / đúng đắn sẽ làm cho việc thống kê đạt kết quả tốt, phản ánh càng chính xác thông tin của population

2.3.Các kỹ thuật lấy mẫu

- Phân loại
  - Các kỹ thuật lấy mẫu được chia làm 2 nhóm chính:
    - Probability sampling (Lấy mẫu xác suất)
      - Lựa chọn mẫu ngẫu nhiên
    - Non-Probability sampling (Lấy mẫu phi xác suất)
      - Lựa chọn mẫu theo ý đồ của người lấy mẫu
- Probability sampling gồm có các kỹ thuật
  - Simple random sampling (Lấy mẫu xác suất đơn giản)
    - Mỗi phần tử của population được chọn với sự ngẫu nhiên như nhau
    - Thực hiện:
      - Các phần tử của population được sắp xếp, gán số thứ tự
      - Tạo ra các giá trị ngẫu nhiên (bốc thăm, dùng phần mềm, …)
      - Chọn phần tử có số thứ tự bằng với số ngẫu nhiên

2.3.Các kỹ thuật lấy mẫu

- Probability sampling gồm có các kỹ thuật
  - Systematic sampling (Lấy mẫu ngẫu nhiên hệ thống)
    - Các phần tử của population được sắp xếp, gán số thứ tự
    - Xác định tổng số các phần tử của population là N
    - Xác định cỡ mẫu cần lấy là n
    - Chia N phần tử thành k nhóm, với $k = N : n$
      (k gọi là khoảng cách chọn mẫu với cỡ mẫu là n)
    - Chọn ngẫu nhiên một nhóm trong k nhóm
    - Chọn nhóm tiếp theo cách nhóm đầu có thể là $2k, 3k, ...$
      cho đến khi đủ số lượng mẫu cần lấy

2.3.Các kỹ thuật lấy mẫu

- Probability sampling gồm có các kỹ thuật
  - Clustering sampling (Lấy mẫu khối)
    - Các phần tử của population được chia thành các khối nhỏ
    - Mỗi khối nhỏ được xem như là một population con
    - Thực hiện phân tích thống kê trên các population con đó
    - Thường dùng khi không thể thu thập đầy đủ population chính
    - VD:
      - Population cần nghiên cứu là các trường đại học
      - Population con là lớp / các lớp

2.3.Các kỹ thuật lấy mẫu

- Probability sampling gồm có các kỹ thuật
  - Stratified sampling (Lấy mẫu phân tầng)
    - Được sử dụng khi các phần tử của population có các đặc điểm quá khác nhau
    - Thực hiện:
      - Xác định độ đo sự tương đồng
      - Các phần tử có độ tương đồng gần giống nhau được xếp chung vào một tầng
      - Áp dụng các kỹ thuật lấy mẫu ở trên (đơn giản, hệ thống, khối) để chọn các mẫu trong các tầng này
    - Vấn đề cần chú ý:
      - Xác định độ đo nào là phù hợp?
      - Cách chia các phần tử vào các tầng → dùng ngưỡng tương đồng là khoảng giá trị nào?

2.3.Các kỹ thuật lấy mẫu

- Non-Probability sampling
  - Các kỹ thuật phi xác suất được dùng trong trường hợp
    - Không sử dụng được các kỹ thuật xác suất, do:
      - Điều kiện thời gian
      - Số lượng
      - Đặc điểm của các phần tử trong population
      - Chi phí thực hiện lấy mẫu ngẫu nhiên
      - . . .
  - Ý nghĩa của các mẫu chọn theo kiểu phi xác suất:
    - Không dùng để làm đại diện cho population
    - Chỉ được chấp nhận để
      - Nghiên cứu khám phá
      - Kiểm định giả thuyết thống kê

2.3.Các kỹ thuật lấy mẫu

- Non-Probability sampling gồm các kỹ thuật
  - Convenient sampling (Lấy mẫu thuận tiện)
    - Mẫu được chọn vì thuận lợi và dễ tiếp cận
    - Tác dụng:
      - Các mẫu này thường dùng để ước lượng sơ bộ vấn đề cần quan tâm đối với population mà không muốn mất quá nhiều thời gian
  - Quota sampling (Lấy mẫu định mức)
    - Người xử lý tùy ý xác định số lượng các phần tử cần quan sát trong các population con
    - Vấn đề:
      - Tỷ lệ số lượng phần tử trong từng population con đối với số lượng phần tử của population bao nhiêu là phù hợp?

2.3.Các kỹ thuật lấy mẫu

- Non-Probability sampling gồm các kỹ thuật
  - Lấy mẫu phán đoán (Judgement sampling)
    - Số lượng mẫu được chọn để quan sát, phân tích được quyết định bởi kinh nghiệm, kiến thức của người xử lý
    - Vấn đề:
      - Chất lượng của mẫu phụ thuộc nhiều vào kinh nghiệm, kiến thức của người lấy mẫu

Q / A