THỐNG KÊ & PHÂN TÍCH DỮ LIỆU

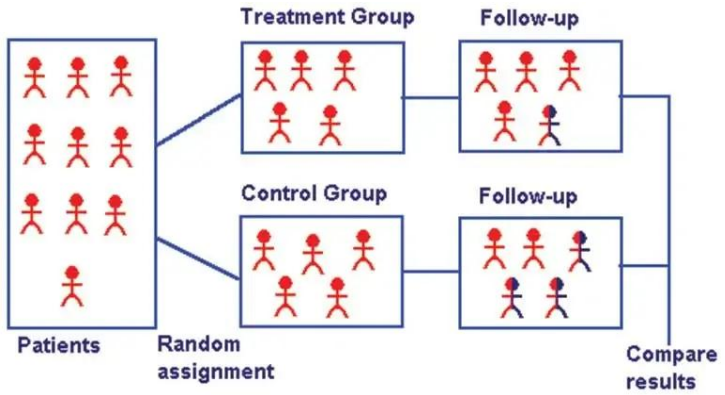
30LT+30TH

GV: TS. Nguyễn Bảo Ân

annb@tvu.edu.vn - 0907966998

Vì sao chúng ta cần thống kê?

- Sự không chắc chắn: Mọi quyết định đều dựa trên những thông tin không đầy đủ
- Ví dụ:
  - Mức lương của sau khi ra trường của một kỹ sư IT là bao nhiêu?
  - Nên chọn kênh đầu tư nào (vàng, đôla, chứng khoán, bất động sản, tiền số, NFT...)?
  - Vaccine A có hiệu quả phòng ngừa cao hơn loại vaccine B không?
  - Tứ hành xung thì sao?
- Các con số và dữ liệu có thể hỗ trợ ra quyết định
- Thống kê là công cụ giúp chúng ta xử lý, tổng hợp, phân tích và giải thích dữ liệu

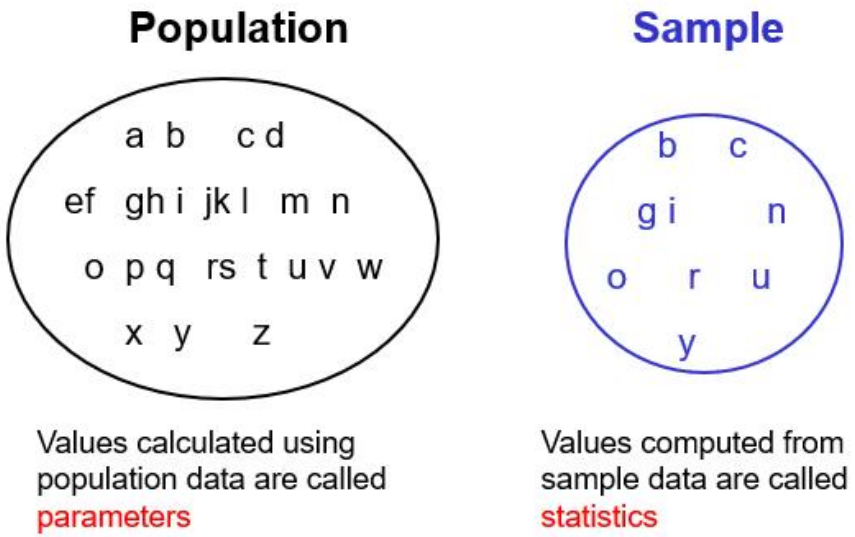


Vì sao chúng ta cần thống kê?

- Nghịch lý Simpson

Một số định nghĩa

- Tổng thể (population): tập hợp tất cả các mẫu mà ta quan tâm
  - N: kích thước/cỡ tổng thể
- Quần thể (sample): là một tập con của tổng thể mà ta quan sát được
  - n: kích thước/cỡ quần thể
- Tham số (parameter): là một tính chất cụ thể của một tổng thể.
- Thống kê (statistics): là một tính chất cụ thể của một quần thể.



Tổng thể - Ví dụ

- Tên của tất cả người dân Việt Nam
- Thu nhập của toàn bộ người dân Trà Vinh
- Điểm trung bình tốt nghiệp của toàn bộ SV khoa Kỹ thuật & Công nghệ
- Cân nặng của toàn bộ SV ngành CNTT cả nước

Lấy mẫu ngẫu nhiên (Random Sampling)

- Trong thực tế, ta muốn biết tham số của tổng thể (VD: chiều cao trung bình của toàn thể SV VN), tuy nhiên, việc lấy số liệu của tổng thể thường không khả thi.
- Lấy mẫu ngẫu nhiên là thủ tục trong đó:
  - Mỗi thành viên của tổng thể được lựa chọn một cách tình cờ
  - Mỗi thành viên của tổng thể đều có khả năng được chọn như nhau
  - Mọi quần thể có thể có của n đối tượng đều có khả năng được chọn như nhau
- Kết quả của lấy mẫu ngẫu nhiên là một quần thể mà ta kỳ vọng có những đặc trưng thống kê giống như của tổng thể.

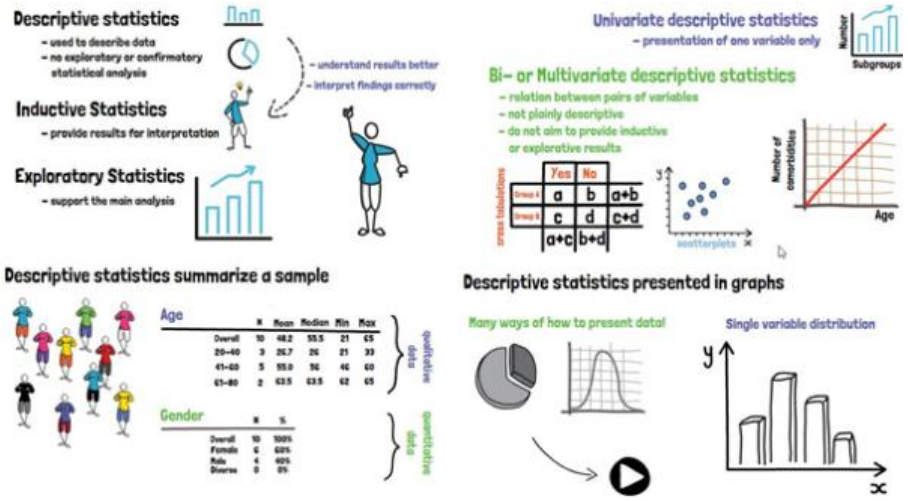
Simple Random Sampling

Source: QuestionPro

Thống kê mô tả và Suy luận thống kê
Descriptive statistics and Inferential Statistics

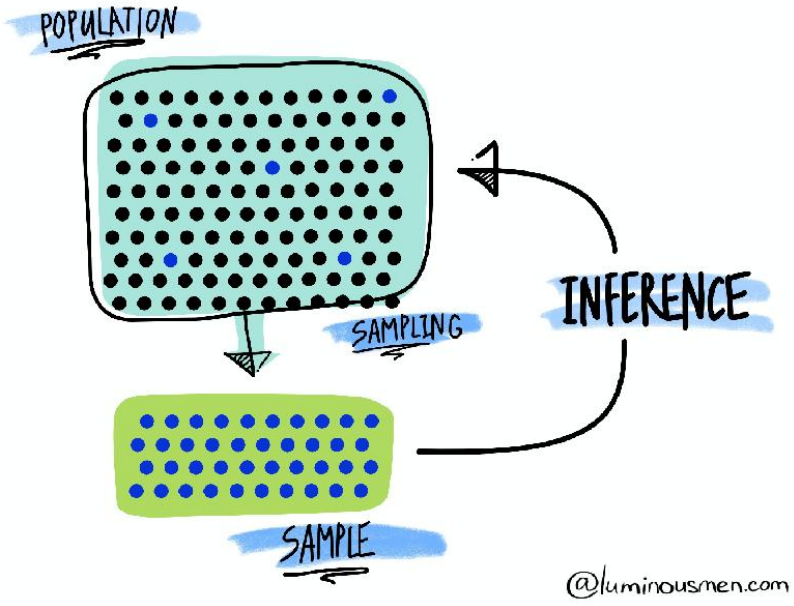
Hai nhánh của thống kê

- Thống kê mô tả
  - Sử dụng số và biểu đồ để tổng kết, mô tả lại dữ liệu



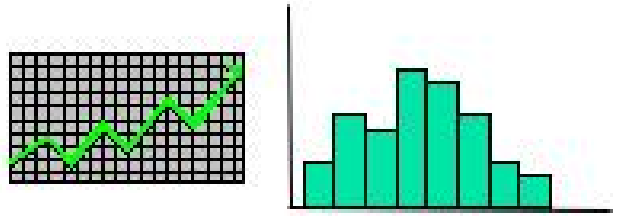
Source: gcp-service.com

- Suy luận thống kê
  - Sử dụng dữ liệu để ra tiên đoán, dự báo, ước lượng để hỗ trợ ra quyết định



Thống kê mô tả

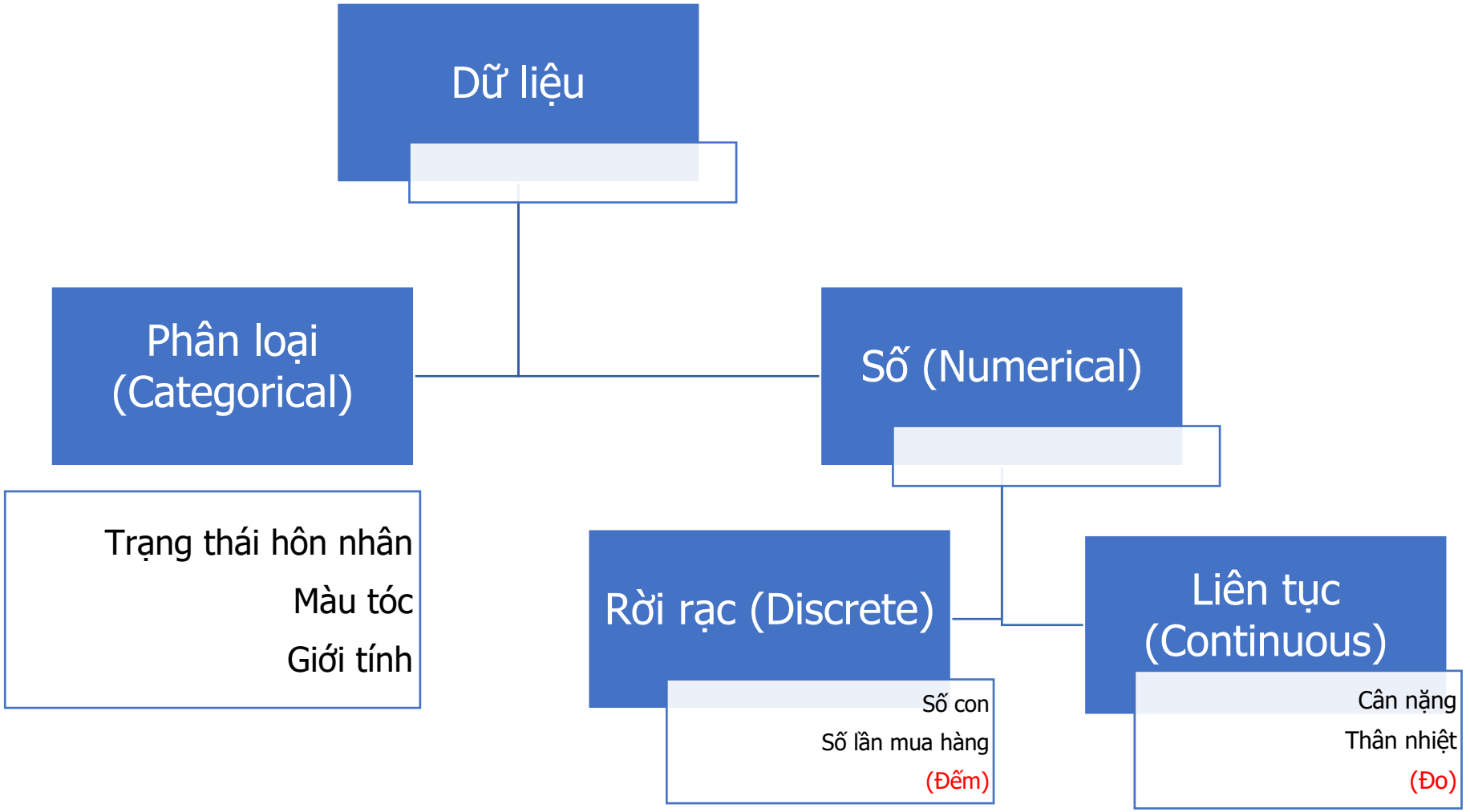
- Thu thập dữ liệu
  - VD: survey
- Trình bày dữ liệu
  - VD: Biểu đồ, bảng biểu
- Tóm tắt dữ liệu
  - VD: Giá trị trung bình, độ lệch chuẩn, biểu đồ tần suất
  - Sample mean = $\frac{\sum X_i}{n}$



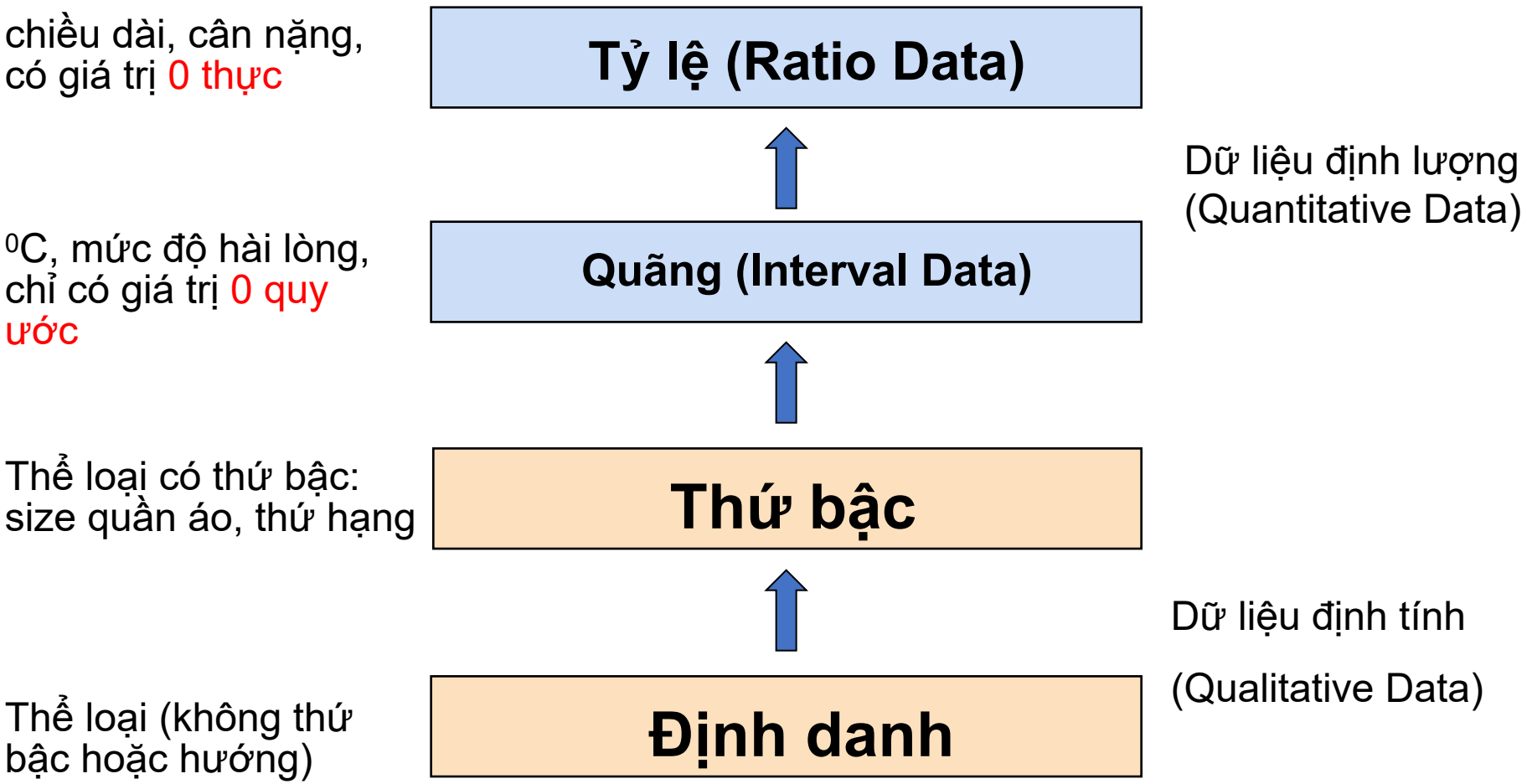
Suy luận thống kê

- Ước lượng:
  - VD: Ước lượng cân nặng trung bình của tổng thể dựa vào trọng lượng trung bình của một quần thể (lấy mẫu ngẫu nhiên)
- Kiểm định giả thuyết
  - VD: Kiểm định giả thuyết:
    - Liệu rằng cân nặng trung bình của tổng thể này là 65.5 kg?
    - Liệu rằng điểm TB môn Kỹ thuật lập trình của sv Nam khác biệt so với sv Nữ? Sự khác biệt này có ý nghĩa thống kê không hay do ngẫu nhiên?
- Suy luận là một quá trình kết luận hoặc ra quyết định cho tổng thể dựa trên kết quả thu được từ quần thể.

Kiểu dữ liệu


Dữ liệu
Phân loại (Categorical)
Trạng thái hôn nhân
Màu tóc
Giới tính
Số (Numerical)
Rời rạc (Discrete)
Số con
Số lần mua hàng
(Đếm)
Liên tục (Continuous)
Cân nặng
Thân nhiệt
(Đo)

Các loại thang đo dữ liệu



Tools...

- Python:
  - Numpy, Pandas, Scipy
- Google Colab
- GitHub
- Get Start