Dữ liệu phân loại và dữ liệu định lượng

Dữ liệu có thể được phân loại thêm thành **dữ liệu phân loại (categorical data)** hoặc **dữ liệu định lượng (quantitative data)**.

- **Dữ liệu phân loại** bao gồm các **nhãn hoặc tên dùng để xác định một thuộc tính của từng phần tử**.
  Chúng sử dụng **thang đo định danh (nominal)** hoặc **thang đo thứ bậc (ordinal)** và có thể **không phải là số** hoặc **được mã hóa bằng số** (ví dụ: 1 = Nam, 2 = Nữ).
- **Dữ liệu định lượng** là dữ liệu yêu cầu **giá trị số biểu thị số lượng hoặc mức độ**, và được thu thập bằng **thang đo khoảng (interval)** hoặc **thang đo tỷ lệ (ratio)**.

Biến phân loại và biến định lượng

- **Biến phân loại (categorical variable)** là biến mà giá trị của nó là dữ liệu phân loại.
- **Biến định lượng (quantitative variable)** là biến có giá trị định lượng.

Việc lựa chọn phương pháp phân tích thống kê phù hợp phụ thuộc vào loại biến: **biến phân loại hay biến định lượng**.

Khi là biến phân loại

- Phân tích thống kê thường **hạn chế hơn**.
- Ta có thể **đếm số lượng quan sát** trong mỗi nhóm hoặc **tính tỷ lệ** phần trăm.
- Ngay cả khi dữ liệu được mã hóa bằng số (ví dụ: 1, 2, 3), các phép toán như cộng, trừ, nhân, chia **không mang ý nghĩa**.

□□ *Ví dụ*: Nếu bạn khảo sát ngành học của 100 sinh viên (Kinh tế, Kế toán, Marketing), thì việc cộng "Kế toán + Marketing" hoàn toàn **không có ý nghĩa gì cả**.

Khi là biến định lượng

- Các phép toán số học **có ý nghĩa thực tiễn**.
- Bạn có thể cộng các giá trị và chia trung bình để ra **giá trị trung bình**, hoặc đo **độ lệch chuẩn**, **phương sai**, v.v.

□□ *Ví dụ*: Bạn có dữ liệu về **thu nhập hàng tháng** của 1.000 người lao động → bạn có thể:

- Tính **thu nhập trung bình**

- Tính **thu nhập tối đa, tối thiểu**
- Vẽ biểu đồ phân phối
- Phân tích xu hướng theo ngành hoặc khu vực

Thực tế trong kinh doanh:

Loại dữ liệu	Ví dụ kinh doanh	Loại biến	Phân tích được áp dụng
Tên sản phẩm	Vinamilk, CocaCola	Phân loại	Đếm số sản phẩm, phân tích tỷ lệ
Ngành hàng	Sữa, Bia, Đồ gia dụng	Phân loại (Ordinal)	Xếp hạng doanh số theo ngành
Doanh thu tháng	12 tỷ, 15 tỷ, 10 tỷ	Định lượng	Trung bình, độ lệch chuẩn, biểu đồ
Mức độ hài lòng (1-5)	1 = rất không hài lòng → 5 = rất hài lòng	Thứ bậc (Ordinal)	Tính trung bình, phân tích xu hướng

Dữ liệu chéo và dữ liệu chuỗi thời gian

Trong phân tích thống kê, việc phân biệt giữa **dữ liệu chéo (cross-sectional data)** và **dữ liệu chuỗi thời gian (time series data)** là rất quan trọng.

- **Dữ liệu chéo** là dữ liệu được thu thập **tại cùng một thời điểm hoặc trong một khoảng thời gian rất ngắn**, từ **nhiều đối tượng khác nhau** (Vinamilk, FPT, Hòa Phát...)
- Ví dụ, bảng dữ liệu dưới đây thể hiện thông tin về khối lượng giao dịch và giá trị giao dịch của 6 công ty niêm yết trên sàn HOSE trong **ngày 1 tháng 4 năm 2025** → Đây là **dữ liệu chéo**.
- **Dữ liệu chuỗi thời gian** là dữ liệu được thu thập **trong nhiều khoảng thời gian liên tiếp** (ví dụ: theo tháng, theo quý, theo năm...).
- Ví dụ: nếu bạn theo dõi giá cổ phiếu VNM từ năm 2020 đến 2025 mỗi tháng → đó là **dữ liệu chuỗi thời gian**.

Công ty	Mã CK	Ngành hàng	KL giao dịch (cổ phiếu)	Giá trị giao dịch (tỷ VNĐ)
Vinamilk	VNM	Sữa & Đồ uống	1,200,000	72.5
FPT	FPT	Công nghệ thông tin	850,000	95.8
Hòa Phát	HPG	Thép & VLXD	2,100,000	102.3
Thế Giới Di Động	MWG	Bán lẻ điện tử	640,000	47.6
Vietcombank	VCB	Ngân hàng	1,750,000	135.2
Sabeco	SAB	Bia & Giải khát	300,000	50.1

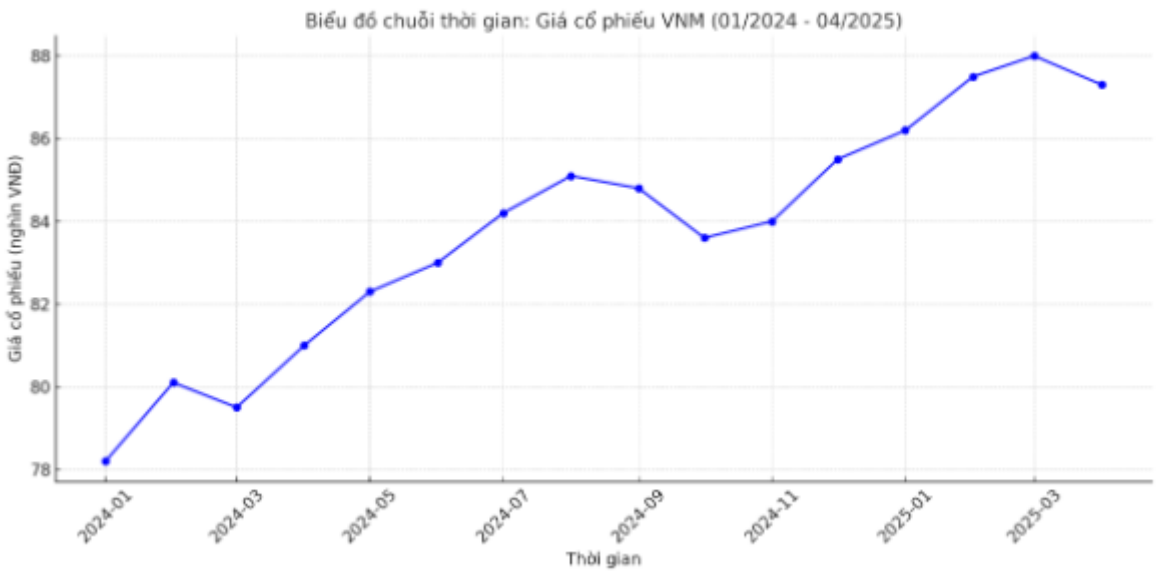
Bảng 1.2 – Dữ liệu chéo minh họa thị trường chứng khoán Việt Nam (01/04/2025)

Phân tích :

- **Dữ liệu định lượng:** Khối lượng giao dịch, Giá trị giao dịch.
- **Dữ liệu phân loại:** Tên công ty, Mã cổ phiếu, Ngành hàng.
- **Thang đo:**
  - Tên công ty, Mã cổ phiếu: **Định danh (Nominal)**.
  - Ngành hàng: **Thứ bậc (Ordinal)** – có thể phân loại theo mức độ ảnh hưởng thị trường.
  - Khối lượng, Giá trị giao dịch: **Tỷ lệ (Ratio)** – có số 0 và đơn vị đo lường có ý nghĩa.

Phân biệt dữ liệu rời rạc và liên tục

- **Dữ liệu rời rạc (discrete):** Là dữ liệu định lượng dùng để **đo đếm số lượng**, ví dụ: **số lượng cổ phiếu giao dịch**, **số lượng nhân viên**.
- **Dữ liệu liên tục (continuous):** Là dữ liệu định lượng dùng để **đo lường**, ví dụ: **giá trị giao dịch (VNĐ)**, **thu nhập**, **trọng lượng hàng hóa** → không có khoảng cách giữa các giá trị liên tiếp.



Biểu đồ chuỗi thời gian thể hiện sự biến động giá cổ phiếu của **VNM (Vinamilk)** từ tháng 1 năm 2024

Biểu đồ này minh họa rõ cách dữ liệu **time series** ghi lại sự thay đổi của một biến số (ở đây là giá cổ phiếu) theo thời gian. Ví dụ như:

- Tháng 1/2024: 78.2 nghìn VNĐ
- Tháng 6/2024: 83.0 nghìn VNĐ

- Tháng 12/2024: 85.5 nghìn VNĐ
- Tháng 4/2025: 87.3 nghìn VNĐ

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Phiên bản #2
Được tạo 24 tháng 4 2025 02:38:38 bởi Đỗ Ngọc Tú
Được cập nhật 24 tháng 4 2025 10:15:13 bởi Đỗ Ngọc Tú