Dữ liệu và Thống kê

Sau khi đọc chương này và hoàn thành các bài tập, bạn sẽ có thể:
- Hiểu và đánh giá được phạm vi ứng dụng rộng lớn của thống kê trong kinh doanh và kinh tế.
- Hiểu ý nghĩa của các thuật ngữ đối tượng nghiên cứu (elements), biến (variables) và quan sát (observations) trong ngữ cảnh thống kê.
- Phân biệt được giữa dữ liệu định tính (qualitative), dữ liệu định lượng (quantitative), dữ liệu theo không gian (cross-sectional) và dữ liệu theo chuỗi thời gian (time series).
- Tìm hiểu về các nguồn dữ liệu phục vụ cho phân tích thống kê, bao gồm cả nguồn nội bộ và bên ngoài doanh nghiệp.
- Nhận thức được cách mà lỗi (errors) có thể phát sinh trong dữ liệu và ảnh hưởng đến kết quả phân tích.
- Hiểu khái niệm thống kê mô tả (descriptive statistics) và suy luận thống kê (statistical inference). Phân biệt được giữa tổng thể (population) và mẫu (sample) trong nghiên cứu thống kê.
- Hiểu vai trò của mẫu trong việc đưa ra các suy luận thống kê về tổng thể.

- Bài tập: Thống kê về quyết định tăng lương
- Bài tập: Thống kê về nguyên nhân tử vong ở Việt Nam
- Bài tập: Thống kê về độc giả tạp chí kinh tế tại Việt Nam

Ứng dụng thống kê trong kinh doanh

Kinh doanh và kinh tế toàn cầu hiện nay, bất kỳ ai cũng có thể tiếp cận một khối lượng lớn thông tin thống kê. Tuy nhiên, những nhà quản lý và người ra quyết định thành công nhất là những người **hiểu rõ dữ liệu thống kê và biết cách vận dụng chúng một cách hiệu quả**.
Trong phần này, chúng ta sẽ xem một số ví dụ minh họa cho việc **thống kê được ứng dụng như thế nào trong lĩnh vực kinh doanh và kinh tế**, từ đó giúp bạn hình dung rõ hơn vai trò quan trọng của thống kê trong việc hỗ trợ phân tích và ra quyết định.

Kế toán (Accounting)

Các công ty kiểm toán độc lập thường sử dụng **các phương pháp chọn mẫu thống kê (statistical sampling)** khi tiến hành kiểm toán cho khách hàng.
Ví dụ, giả sử một công ty kiểm toán muốn xác định liệu số dư **phải thu khách hàng (accounts receivable)** được trình bày trên bảng cân đối kế toán của khách hàng có phản ánh trung thực giá trị thực tế hay không. Trong thực tế, số lượng các khoản phải thu thường rất lớn, nên việc kiểm tra từng khoản một sẽ tốn quá nhiều thời gian và chi phí.

Do đó, một thông lệ phổ biến là nhóm kiểm toán sẽ chọn **một tập hợp con của các khoản phải thu**, gọi là **mẫu (sample)**. Sau khi kiểm tra độ chính xác của các khoản trong mẫu, kiểm toán viên sẽ đưa ra kết luận liệu con số phải thu được trình bày trong báo cáo tài chính có hợp lý và chấp nhận được hay không.

Phương pháp chọn mẫu thống kê trong kiểm toán không chỉ giúp tiết kiệm nguồn lực mà còn tạo điều kiện để **áp dụng các kỹ thuật suy luận thống kê (statistical inference)** vào việc đưa ra nhận định cho cả tổng thể.

Một điều quan trọng trong quá trình này là **lựa chọn mẫu ngẫu nhiên và đại diện**, nhằm đảm bảo rằng kết quả từ mẫu có thể suy rộng đáng tin cậy cho toàn bộ dữ liệu. Ngoài ra, các kiểm toán viên còn phải xác định mức **sai số chấp nhận được (margin of error)** và **mức độ tin cậy (confidence level)** để làm căn cứ cho các kết luận của mình.

Tài chính (Finance)

Các chuyên gia phân tích tài chính sử dụng nhiều loại thông tin thống kê khác nhau để đưa ra khuyến nghị đầu tư.
Trong trường hợp cổ phiếu, họ thường xem xét nhiều dữ liệu tài chính như **tỷ số giá trên thu**

nhập (P/E - price/earnings ratio) và **lợi suất cổ tức (dividend yield)**.

> **Tỷ số P/E** = Giá thị trường của cổ phiếu / Lợi nhuận trên mỗi cổ phiếu (EPS)
>
> Nó thể hiện **nhà đầu tư sẵn sàng trả bao nhiêu tiền cho 1 đồng lợi nhuận mà công ty tạo ra**.
> Ví dụ: P/E = 15 nghĩa là nhà đầu tư trả 15 đồng để thu về 1 đồng lợi nhuận.

Bằng cách so sánh thông tin của một cổ phiếu cụ thể với các chỉ số trung bình của thị trường chứng khoán, nhà phân tích tài chính có thể đưa ra nhận định liệu cổ phiếu đó đang **được định giá quá cao (overpriced)** hay **quá thấp (underpriced)**.

Tương tự, các **xu hướng giá cổ phiếu trong quá khứ (historical price trends)** cũng có thể cung cấp những chỉ báo quan trọng giúp nhà đầu tư xác định thời điểm nên tham gia hoặc quay lại thị trường.

Ví dụ, vào ngày 3 tháng 4 năm 2009, tạp chí *Money Week* đưa tin về một phân tích của Goldman Sachs cho rằng, do giá cổ phiếu lúc đó đang ở mức đặc biệt thấp, nhà đầu tư có thể kỳ vọng mức lợi nhuận trung bình thực tế lên tới **6% tại Hoa Kỳ** và **7% tại Vương quốc Anh** trong vòng một thập kỷ tới – dựa trên **tỷ lệ P/E được điều chỉnh theo chu kỳ dài hạn**.

Lĩnh vực tài chính là một trong những môi trường **ứng dụng thống kê mạnh mẽ và rộng rãi nhất**. Các nhà phân tích không chỉ sử dụng thống kê mô tả để tóm tắt dữ liệu, mà còn thường xuyên sử dụng các mô hình thống kê suy luận và mô hình dự báo (forecasting models), chẳng hạn như **hồi quy tuyến tính, phân tích chuỗi thời gian, và mô hình ARIMA**.

Thống kê giúp giảm thiểu rủi ro trong các quyết định tài chính bằng cách cung cấp một nền tảng dữ liệu có hệ thống, giúp nhà đầu tư **không ra quyết định dựa trên cảm tính**, mà dựa trên các chỉ số và mô hình có cơ sở khoa học.

Tiếp thị (Marketing)

Các thiết bị quét mã vạch (electronic scanners) tại quầy thanh toán của các cửa hàng bán lẻ **thu thập dữ liệu phục vụ cho nhiều mục đích nghiên cứu thị trường khác nhau**.

Ví dụ, các nhà cung cấp dữ liệu như **ACNielsen** mua lại dữ liệu từ máy quét tại điểm bán (point-of-sale scanner data) từ các cửa hàng tạp hóa, xử lý dữ liệu đó, và sau đó **bán lại các bảng tổng hợp thống kê** cho các nhà sản xuất.
Các nhà sản xuất thường chi một khoản tiền lớn cho từng nhóm sản phẩm để sở hữu loại dữ liệu này.

Ngoài ra, họ còn mua dữ liệu và các bản tổng hợp thống kê liên quan đến **hoạt động xúc tiến bán hàng (promotional activities)** như:

– **Chương trình giảm giá đặc biệt (special pricing)**
– **Trưng bày sản phẩm trong cửa hàng (in-store displays)**

Các **quản lý thương hiệu (brand managers)** có thể phân tích **thống kê từ dữ liệu máy quét** và **thống kê từ hoạt động khuyến mãi** để hiểu rõ hơn **mối quan hệ giữa các chương trình khuyến mãi và doanh số bán hàng**.

Những phân tích này thường **cung cấp thông tin có giá trị** để xây dựng **chiến lược tiếp thị hiệu quả hơn cho từng sản phẩm trong tương lai**.

Đây là ví dụ điển hình về cách **thống kê giúp chuyển đổi dữ liệu thô thành tri thức chiến lược**.
Thông qua phân tích mối quan hệ giữa **biến số khuyến mãi** và **kết quả bán hàng**, các nhà tiếp thị có thể:

- **Dự đoán hiệu quả của các chương trình khuyến mãi**,
- **Tối ưu hóa ngân sách marketing**,
- **Cá nhân hóa chiến lược theo từng nhóm khách hàng**.

Ngày nay, với sự phát triển của **phân tích dữ liệu lớn (big data analytics)** và **AI trong marketing**, vai trò của thống kê càng trở nên quan trọng và sâu sắc hơn bao giờ hết.

Sản xuất (Production)

Ngày nay, khi **chất lượng được đặt lên hàng đầu**, thì **kiểm soát chất lượng (quality control)** trở thành một ứng dụng quan trọng của thống kê trong lĩnh vực sản xuất.

Nhiều loại **biểu đồ kiểm soát thống kê (statistical quality control charts)** được sử dụng để **giám sát đầu ra của quy trình sản xuất**.
Trong đó, **biểu đồ trung bình $\bar{x}$ (x-bar chart)** là một công cụ phổ biến để theo dõi **giá trị trung bình của sản phẩm đầu ra**.

Ví dụ, giả sử một **máy chiết rót nước giải khát** được lập trình để rót **330g** mỗi chai.
Định kỳ, công nhân sản xuất sẽ **lấy một mẫu ngẫu nhiên** các chai và tính **trung bình lượng nước ($\bar{x}$)** trong mẫu.
Giá trị trung bình này được **vẽ lên biểu đồ x-bar**.

- Nếu giá trị trung bình **vượt quá giới hạn trên (Upper Control Limit - UCL)** của biểu đồ, điều đó cho thấy **máy đang rót quá mức (overfilling)**.
- Nếu giá trị **thấp hơn giới hạn dưới (Lower Control Limit - LCL)**, tức là **rót thiếu (underfilling)**.

Quy trình được coi là **"đang trong kiểm soát" (in control)** và **có thể tiếp tục** nếu các điểm dữ liệu nằm **giữa hai giới hạn kiểm soát**.

Khi được phân tích đúng cách, **biểu đồ x̄ giúp xác định thời điểm cần điều chỉnh máy móc hoặc quy trình** để đảm bảo chất lượng sản phẩm luôn đạt tiêu chuẩn.

rong ngành sản xuất **sữa tiệt trùng**, một nhà máy đóng gói sữa hộp 1 lít.
Cứ sau mỗi 30 phút, kỹ thuật viên lấy ngẫu nhiên **5 hộp sữa**, cân từng hộp, và tính trung bình.
Nếu trung bình nằm ngoài giới hạn cho phép (ví dụ dưới 980ml hoặc trên 1020ml), kỹ thuật viên **ngay lập tức điều chỉnh máy chiết rót**.

Phân tích thống kê này giúp **giảm tỷ lệ hàng lỗi, tiết kiệm chi phí sản xuất,** và **duy trì uy tín thương hiệu**.

Kinh tế học (Economics)

Các nhà kinh tế học thường **đưa ra dự báo về tương lai của nền kinh tế hoặc một khía cạnh cụ thể nào đó của nó**.
Để xây dựng các dự báo này, họ sử dụng **nhiều loại dữ liệu thống kê khác nhau**.

Ví dụ, trong việc **dự báo tỷ lệ lạm phát**, các nhà kinh tế thường sử dụng các **chỉ số thống kê** như:

- **Chỉ số giá sản xuất (Producer Price Index – PPI)**
- **Tỷ lệ thất nghiệp (Unemployment rate)**
- **Mức sử dụng công suất sản xuất (Manufacturing capacity utilization)**

Thông thường, các chỉ số thống kê này được đưa vào **các mô hình dự báo được máy tính hóa**, từ đó **tạo ra các ước tính về tỷ lệ lạm phát trong tương lai**.

Chẳng hạn, **Ngân hàng Trung ương Việt Nam (SBV)** có thể dựa vào dữ liệu **PPI, CPI, tăng trưởng GDP và tỷ lệ thất nghiệp** để:

- **Dự báo xu hướng lạm phát quý sau**
- Quyết định **điều chỉnh lãi suất điều hành** để kiểm soát lạm phát và ổn định kinh tế vĩ mô

Một ví dụ khác, **Quỹ Tiền tệ Quốc tế (IMF)** và **Ngân hàng Thế giới (WB)** thường sử dụng **mô hình kinh tế lượng (econometric models)** để dự báo **mức tăng trưởng của các nền kinh tế đang phát triển**, trong đó Việt Nam là một trường hợp điển hình.

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

DỮ LIỆU (DATA)

I. Dữ liệu

Dữ liệu là **các sự kiện và con số được thu thập, phân tích và tóm tắt** nhằm phục vụ cho **việc trình bày và diễn giải**.

Toàn bộ dữ liệu được thu thập trong một nghiên cứu cụ thể được gọi là **tập dữ liệu (data set)** của nghiên cứu đó.

Company	Ticker	Sector	Volume Traded (shares)	Price (VND)	Market Capitalization (Billion VND)	Price Change (%)	Date
VinGroup	VIC	Real Estate	2,500,000	100,000	250,000	1.5	2024-03-01
Vietcombank	VCB	Banking	3,500,000	120,000	420,000	-0.5	2024-03-01
HoaPhat	HPG	Steel	4,200,000	55,000	300,000	0.3	2024-03-01
Masangroup	MSN	Food & Beverage	3,000,000	150,000	120,000	2.0	2024-03-01
BIDV	BID	Banking	2,800,000	42,000	220,000	-1.2	2024-03-01

Dưới đây là **Bảng 1.1** minh họa **dữ liệu giao dịch cổ phiếu** tại **Sở Giao dịch Chứng khoán Việt Nam** (VNX) cho tháng 3 năm 2024:

Giải thích:

- **Volume Traded (shares)**: Số lượng cổ phiếu giao dịch.
- **Ticker:** Mã cổ phiếu
- **Sector:** Ngành
- **Price (VND)**: Giá cổ phiếu tại thời điểm giao dịch.
- **Market Capitalization (Billion VND)**: Vốn hóa thị trường của công ty (theo tỷ lệ cổ phiếu phát hành và giá cổ phiếu).
- **Price Change (%)**: Tỷ lệ thay đổi giá cổ phiếu so với ngày hôm trước.
- **Date**: Ngày giao dịch.

Theo báo cáo cập nhật từ **Sở Giao dịch Chứng khoán London (LSE)** tháng 3/2024, tổng khối lượng giao dịch cổ phiếu đạt **hơn 1.1 nghìn tỷ GBP**, với **các ngành năng lượng và AI tăng trưởng mạnh nhất**.

Tập dữ liệu được sử dụng trong báo cáo này bao gồm:

- **Tên công ty niêm yết**
- **Tổng giá trị giao dịch trong tháng**
- **Biến động giá cổ phiếu**
- **Tỷ lệ tăng trưởng so với tháng trước**

> Những tập dữ liệu như vậy đang ngày càng trở nên phổ biến nhờ sự phát triển của **dữ liệu lớn (big data)** và **AI phân tích dữ liệu (data analytics AI)**.

II. Các yếu tố (Elements), Biến số (Variables), Quan sát (Observations)

Các yếu tố (Elements): Là các thực thể mà dữ liệu được thu thập trên đó. Trong bảng dữ liệu này, mỗi công ty niêm yết là một yếu tố (VinGroup, Vietcombank, HoaPhat...). Có 5 công ty nên bộ dữ liệu chứa 5 yếu tố.

Biến số (Variables): Là các đặc tính quan tâm của các yếu tố. Bảng này gồm 7 biến:

- Công ty (Company)
- Mã chứng khoán (Ticker)
- Ngành (Sector)
- Khối lượng giao dịch (Volume Traded - shares)
- Giá (Price - VND)
- Vốn hóa thị trường (Market Capitalization - Billion VND)
- Thay đổi giá (%) (Price Change)
- Ngày (Date)

Quan sát (Observations): Là tập hợp các giá trị đo lường thu thập được cho một yếu tố cụ thể. Mỗi hàng trong bảng là một quan sát. Ví dụ quan sát đầu tiên (VinGroup) có các giá trị: VIC, Real Estate, 2,500,000, 100,000, 250,000, 1.5%, 2024-03-01.

Phân tích bảng dữ liệu

1. **Các yếu tố (Elements)**: 5 công ty niêm yết trên sàn chứng khoán Việt Nam
   - VinGroup (VIC)
   - Vietcombank (VCB)
   - HoaPhat (HPG)
   - Masangroup (MSN)
   - BIDV (BID)
2. **Các biến số (Variables)**:
   - **Biến định tính (Qualitative)**:
     - Company: Tên công ty

    - Ticker: Mã chứng khoán
    - Sector: Ngành hoạt động
    - Date: Ngày giao dịch
  - **Biến định lượng (Quantitative)**:
    - Volume Traded: Khối lượng cổ phiếu giao dịch (liên tục)
    - Price: Giá cổ phiếu (liên tục)
    - Market Capitalization: Vốn hóa thị trường (liên tục)
    - Price Change: % thay đổi giá (liên tục)
3. **Quan sát (Observations)**: 5 quan sát tương ứng với 5 công ty
4. **Phân tích ngành**:
   - 2 công ty ngành ngân hàng (VCB, BID)
   - 1 công ty bất động sản (VIC)
   - 1 công ty thép (HPG)
   - 1 công ty thực phẩm & đồ uống (MSN)
5. **Giá trị nổi bật**:
   - Giá cao nhất: MSN (150,000 VND)
   - Giá thấp nhất: BID (42,000 VND)
   - Khối lượng giao dịch lớn nhất: HPG (4,200,000 shares)
   - Vốn hóa lớn nhất: VCB (420,000 tỷ VND)
   - Tăng giá mạnh nhất: MSN (+2.0%)
   - Giảm giá nhiều nhất: BID (-1.2%)

III. Thang đo trong thống kê và phân tích dữ liệu

Việc thu thập dữ liệu đòi hỏi phải xác định thang đo phù hợp, bao gồm: định danh (nominal), thứ bậc (ordinal), khoảng cách (interval) hoặc tỷ lệ (ratio). Thang đo quyết định lượng thông tin chứa trong dữ liệu và giúp lựa chọn phương pháp tổng hợp, phân tích thống kê phù hợp.

1. Thang đo định danh (Nominal Scale)

- **Định nghĩa**: Dùng để phân loại dữ liệu dựa trên **nhãn hoặc tên gọi**, không có thứ tự hay giá trị số.
- **Đặc điểm**:
  - Chỉ phân biệt các nhóm, không so sánh hơn/kém.
  - Có thể dùng mã số thay cho nhãn (ví dụ: 1 = VinGroup, 2 = Vietcombank), nhưng con số không mang ý nghĩa toán học.
- **Ví dụ trong bảng 1.1**:
  - *Company* (tên công ty), *Ticker* (mã cổ phiếu), *Sector* (ngành) – đều là biến **định danh**.
  - Ví dụ: *Sector* gồm "Real Estate", "Banking", "Steel"... chỉ phân loại, không xếp hạng.

2. Thang đo thứ bậc (Ordinal Scale)

- **Định nghĩa**: Dữ liệu có thể **sắp xếp theo thứ tự**, nhưng khoảng cách giữa các hạng mục không đồng nhất.
- **Đặc điểm**:

- Có thể dùng số để mã hóa (ví dụ: 1 = Rất tốt, 2 = Tốt, 3 = Kém), nhưng phép trừ/chia giữa các số vô nghĩa.
- Đánh giá dịch vụ: "Xuất sắc", "Khá", "Trung bình".
- Nếu có biến như *Xếp hạng rủi ro* (AAA, AA, B), đó là thang **thứ bậc**.

- **Ví dụ**

Rank	Company	...
1	VinGroup	
2	Vietcombank	
3	HoaPhat	

`Rank` – là thứ hạng của công ty theo tổng giá trị giao dịch (Total Trading Value).
Mặc dù `VinGroup` xếp hạng 1 và `Vietcombank` xếp hạng 2, chúng ta không biết **VinGroup vượt bao nhiêu phần trăm** so với Vietcombank – chỉ biết là đứng **cao hơn**.

3. Thang đo khoảng cách (Interval Scale)

- **Định nghĩa**: Dữ liệu có thứ tự **và khoảng cách giữa các giá trị có ý nghĩa**, nhưng không có điểm "0 tuyệt đối".
- **Đặc điểm**:
  - Phép cộng/trừ có nghĩa, nhưng phép nhân/chia không hợp lý.
  - Ví dụ: Nhiệt độ (°C), điểm GMAT.
- **Ví dụ trong bảng**:
  - *Price Change (%)* (thay đổi giá): Có thể tính chênh lệch (ví dụ: MSN tăng 2.0%, BID giảm 1.2% → khoảng cách MSN và BID là 3.2%), nhưng không thể nói "tăng gấp đôi".

4. Thang đo tỷ lệ (Ratio Scale)

- **Định nghĩa**: Có đủ tính chất của thang khoảng cách **và có điểm 0 tuyệt đối**, cho phép so sánh tỷ lệ.
- **Đặc điểm**:
  - Có thể dùng mọi phép toán (+, -, ×, ÷).
  - Ví dụ: Chiều cao, cân nặng, doanh thu.
- **Ví dụ trong bảng**:
  - *Price* (giá cổ phiếu): 100,000 VND gấp đôi 50,000 VND.
  - *Volume Traded* (khối lượng giao dịch): 2,500,000 cổ phiếu có thể so sánh tỷ lệ với 3,500,000 cổ phiếu.
  - *Market Cap* (vốn hóa): 250,000 tỷ VND = 2.5 lần 100,000 tỷ VND.

Lưu ý quan trọng

- Thang đo **tỷ lệ (ratio)** là mạnh nhất, cho phép sử dụng mọi phương pháp thống kê.
- Biến **định danh** và **thứ bậc** thường dùng cho phân loại hoặc kiểm định phi tham số (ví dụ: Kiểm định Chi-square).

- Hiểu đúng thang đo giúp chọn **đồ thị** phù hợp (ví dụ: biểu đồ cột cho nominal, biểu đồ xếp hạng cho ordinal).

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

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

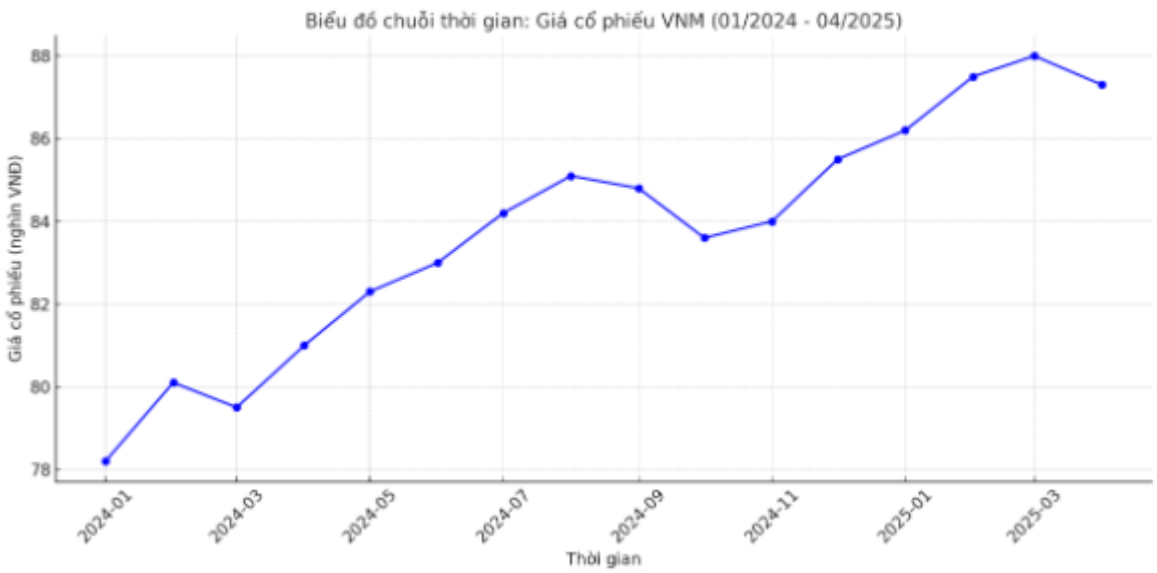
Bảng 1.2 – Dữ liệu chéo minh họa thị trường chứng khoán Việt Nam (01/04/2025)

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

Nguần dữ liệu(DATA SOURCES)

Nguồn dữ liệu có thể đến từ các nguồn sẵn có hoặc được thu thập mới thông qua khảo sát và nghiên cứu thực nghiệm.

Nguồn dữ liệu sẵn có

Trong một số trường hợp, dữ liệu cần thiết cho một ứng dụng cụ thể **đã tồn tại**. Các công ty tại Việt Nam hiện nay thường lưu trữ nhiều cơ sở dữ liệu khác nhau về **nhân viên, khách hàng và hoạt động kinh doanh**. Dữ liệu về **mức lương, độ tuổi và số năm kinh nghiệm của nhân viên** thường có thể lấy từ hồ sơ nhân sự nội bộ. Các hồ sơ nội bộ khác có thể chứa thông tin về:

- Doanh số bán hàng
- Chi phí quảng cáo
- Chi phí phân phối
- Mức tồn kho
- Sản lượng sản xuất

Ngoài ra, **nhiều doanh nghiệp cũng quản lý dữ liệu rất chi tiết về khách hàng** như hành vi mua hàng, tần suất giao dịch, khu vực sinh sống v.v...

Nhóm dữ liệu	Ví dụ dữ liệu có thể thu thập
Nhân sự	Mức lương, chức vụ, số năm làm việc
Khách hàng	Tên, độ tuổi, khu vực, lịch sử mua hàng
Bán hàng	Doanh số theo quý, sản phẩm bán chạy
Quảng cáo	Chi phí chạy quảng cáo trên Facebook, Google
Chuỗi cung ứng	Chi phí vận chuyển, số lượng tồn kho, nhà cung cấp
Sản xuất	Sản lượng, tỉ lệ lỗi sản phẩm, thời gian hoàn thành

Bảng 1.3: Một số dữ liệu thường có trong hệ thống nội bộ của công ty (Việt Nam)

Các nguồn dữ liệu bên ngoài tại Việt Nam

Nhiều tổ chức trong và ngoài nước chuyên thu thập và cung cấp dữ liệu về **kinh tế, thị trường và doanh nghiệp** tại Việt Nam. Một số nguồn thông tin phổ biến:

- **Tổng cục Thống kê Việt Nam (GSO)**: Dữ liệu dân số, thu nhập, lao động, xuất nhập khẩu.
- **Cục Quản lý đăng ký kinh doanh (Bộ KH&ĐT)**: Thông tin về số lượng doanh nghiệp thành lập, giải thể.
- **Cục Thuế, Tổng cục Hải quan**: Dữ liệu về thuế, kim ngạch xuất nhập khẩu.
- **VNDirect, SSI Research**: Cung cấp báo cáo ngành và thông tin chứng khoán.
- **DataViet, InfoTV, Vietdata**: Các công ty chuyên cung cấp dữ liệu thương mại, thị trường tiêu dùng.

Doanh nghiệp có thể tiếp cận các nguồn này **thông qua thuê bao, mua dữ liệu hoặc qua nền tảng mở** của cơ quan nhà nước.

Internet – nguồn dữ liệu ngày càng quan trọng

Ngày nay, **Internet** trở thành một **kho dữ liệu khổng lồ**. Hầu hết các công ty tại Việt Nam đều có website cung cấp:

- Thông tin công ty
- Sản phẩm/dịch vụ
- Giá bán, chương trình khuyến mãi
- Tuyển dụng, quy mô nhân sự
- Tin tức nội bộ, báo cáo tài chính

Ngoài ra, nhiều nền tảng chuyên biệt như:

- **CafeF, Vietstock, Investing.vn**: Cập nhật liên tục thông tin về thị trường tài chính.
- **Google Trends, Facebook Audience Insights**: Dữ liệu xu hướng người dùng.
- **Foody, Shopee, Tiki**: Thống kê đánh giá sản phẩm, giá cả và thói quen mua sắm.

Cơ quan nhà nước – nguồn dữ liệu đáng tin cậy

Các cơ quan nhà nước cũng là nguồn cung cấp dữ liệu rất giá trị. Ví dụ:

Cơ quan	Dữ liệu cung cấp
Tổng cục Thống kê (GSO)	Dân số, GDP, tỉ lệ thất nghiệp, chỉ số giá tiêu dùng
Bộ Giáo dục & Đào tạo	Số lượng sinh viên, điểm thi trung bình, phân tích ngành học
Ngân hàng Nhà nước	Lãi suất, tỷ giá, báo cáo tiền tệ, tín dụng
Tổng cục Hải quan	Kim ngạch xuất – nhập khẩu, đối tác thương mại chủ yếu
Bộ Y tế	Số ca bệnh, cơ sở khám chữa bệnh, phân bổ nhân lực y tế

Hầu hết dữ liệu đều có thể **tải về miễn phí tại các cổng thông tin điện tử** như:

https://www.gso.gov.vn

https://data.gov.vn

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Nghiên cứu thống kê

Đôi khi dữ liệu cần thiết cho một ứng dụng cụ thể không có sẵn từ các nguồn hiện tại. Trong những trường hợp như vậy, dữ liệu có thể được thu thập thông qua một nghiên cứu thống kê. Các nghiên cứu thống kê có thể được phân loại thành **nghiên cứu thực nghiệm** hoặc **nghiên cứu quan sát**.

I. Nghiên cứu thực nghiệm

Trong một nghiên cứu thực nghiệm, trước tiên cần xác định biến cần quan tâm. Sau đó, một hoặc nhiều biến khác sẽ được kiểm soát để thu thập dữ liệu về cách các biến đó ảnh hưởng đến biến cần quan tâm.

Ví dụ, một công ty dược phẩm có thể muốn tiến hành một thí nghiệm để tìm hiểu tác động của một loại thuốc mới đến huyết áp. Huyết áp là biến được quan tâm. Liều lượng của thuốc mới là biến được kỳ vọng có ảnh hưởng đến huyết áp. Để thu thập dữ liệu, các nhà nghiên cứu chọn một mẫu người tham gia và chia thành nhiều nhóm nhận các liều lượng khác nhau. Dữ liệu về huyết áp trước và sau khi dùng thuốc sẽ được thu thập cho mỗi nhóm. Phân tích thống kê dữ liệu thực nghiệm sẽ giúp xác định ảnh hưởng thực sự của thuốc đến huyết áp.

Tình hình tại Việt Nam:

Các nghiên cứu thực nghiệm hiện nay thường được thực hiện tại các bệnh viện lớn như Bệnh viện Chợ Rẫy, Bạch Mai, hoặc tại các trường đại học như Đại học Y Dược TP.HCM, trong các đề tài nghiên cứu thuốc mới, thực phẩm chức năng hoặc các liệu pháp điều trị mới.

Nghiên cứu quan sát (không thực nghiệm)

Nghiên cứu thống kê không thực nghiệm hay còn gọi là **nghiên cứu quan sát** không cố gắng kiểm soát các biến. **Khảo sát** là dạng nghiên cứu quan sát phổ biến nhất. Ví dụ, trong một khảo sát phỏng vấn cá nhân, các câu hỏi nghiên cứu được xác định trước, sau đó thiết kế một bảng câu hỏi và thực hiện với một mẫu người tham gia.

Một số nhà hàng sử dụng khảo sát quan sát để thu thập ý kiến khách hàng về chất lượng món ăn, dịch vụ, không gian, v.v. Một bảng khảo sát tại nhà hàng **Lobster Pot ở thành phố Limerick, Ireland** yêu cầu khách hàng đánh giá 5 tiêu chí: chất lượng món ăn, thái độ phục vụ, thời gian phục vụ, vệ sinh và cách quản lý. Các mức đánh giá gồm: xuất sắc, tốt, đạt yêu cầu và không đạt — được gọi là dữ liệu **xếp hạng (ordinal data)**, giúp nhà quản lý đánh giá hoạt động của nhà hàng.

Tại Việt Nam, các chuỗi như Highlands Coffee, The Coffee House, hoặc nhà hàng Gogi House thường đính kèm mã QR khảo sát trên hóa đơn để thu thập ý kiến khách hàng. Ngoài ra, các doanh nghiệp cũng sử dụng khảo sát online qua Google Forms hoặc các nền tảng như Zoho, SurveyMonkey để nghiên cứu thị trường hoặc đánh giá mức độ hài lòng của khách hàng.

Chi phí và thời gian thu thập dữ liệu

Các nhà quản lý cần hiểu rõ về **thời gian và chi phí** liên quan khi thu thập dữ liệu. Sử dụng các nguồn dữ liệu có sẵn sẽ thuận tiện hơn nếu cần dữ liệu trong thời gian ngắn. Nếu dữ liệu quan trọng không có sẵn, cần xem xét kỹ chi phí và thời gian để thu thập chúng. Dù sao, việc ra quyết định nên dựa trên phân tích thống kê **hiệu quả về chi phí** — tức là chi phí thu thập và phân tích dữ liệu không nên vượt quá lợi ích thu được từ quyết định cải thiện.

II. Sai sót trong thu thập dữ liệu

Các nhà quản lý cũng cần chú ý đến **khả năng sai sót** trong quá trình thu thập dữ liệu. Việc sử dụng dữ liệu sai còn **nguy hiểm hơn** việc không có dữ liệu.

Sai sót có thể xảy ra khi:

- Người thu thập dữ liệu ghi nhầm (ví dụ: ghi tuổi 24 thành 42).
- Người trả lời hiểu sai câu hỏi và đưa ra câu trả lời không chính xác.

Những nhà phân tích dữ liệu có kinh nghiệm thường **rất cẩn trọng** trong việc thu thập và ghi chép dữ liệu. Họ sử dụng các kỹ thuật kiểm tra độ nhất quán nội bộ, ví dụ: nếu một người khai 22 tuổi nhưng có 20 năm kinh nghiệm làm việc thì cần kiểm tra lại dữ liệu. Ngoài ra, họ cũng xem xét các giá trị bất thường (gọi là **outliers**) để loại trừ khả năng sai sót.

Tại Việt Nam, trong các khảo sát trực tiếp tại sự kiện hay qua điện thoại, sai sót thường gặp do người ghi phiếu vội vàng, người trả lời không hợp tác hoặc hiểu sai câu hỏi. Do đó, nên tổ chức các buổi huấn luyện cho người khảo sát và kiểm tra lại phiếu sau khi thu thập để đảm bảo độ chính xác.

III. Mẫu khảo sát thực tế

1. Mẫu khảo sát khách hàng cho Quán Cà Phê

Tiêu đề:

> **Khảo sát trải nghiệm khách hàng tại [Tên quán cà phê]**

Lời mở đầu:

>

> Chúng tôi rất mong nhận được góp ý của bạn để cải thiện chất lượng dịch vụ. Khảo sát chỉ mất khoảng 1 phút.

Câu hỏi:

1. Bạn đến quán vào thời điểm nào trong ngày?
   - Sáng
   - Chiều
   - Tối
2. Bạn đánh giá chất lượng đồ uống như thế nào?
   - Xuất sắc
   - Tốt
   - Bình thường
   - Kém
3. Không gian quán có phù hợp với bạn không?
   - Rất phù hợp
   - Tạm được
   - Không phù hợp
4. Thái độ phục vụ của nhân viên:
   - Thân thiện, nhiệt tình
   - Bình thường
   - Cần cải thiện
5. Bạn có quay lại quán trong tương lai không?
   - Có
   - Không chắc
   - Không
6. Góp ý thêm (tùy chọn):
   ➤ .....

2. Mẫu khảo sát chất lượng dịch vụ Nhà Hàng

Tiêu đề:

> **Phiếu khảo sát chất lượng tại nhà hàng [Tên nhà hàng]**

Câu hỏi:

1. Món ăn được phục vụ có đúng món và đúng yêu cầu không?
   - Có
   - Không
2. Thời gian chờ món:
   - Nhanh
   - Trung bình

- Quá lâu

3. Chất lượng món ăn:
   - Rất ngon
   - Ngon
   - Bình thường
   - Kém
4. Bạn đánh giá thế nào về thái độ nhân viên?
   - Thân thiện, chuyên nghiệp
   - Bình thường
   - Không hài lòng
5. Bạn có giới thiệu nhà hàng cho bạn bè/đồng nghiệp không?
   - Có
   - Không
6. Góp ý khác (nếu có):
   ➤ .....

---

3. Mẫu khảo sát khách hàng cửa hàng bán lẻ (shop, siêu thị mini)

Tiêu đề:

> **Khảo sát hài lòng khách hàng – [Tên cửa hàng]**

Câu hỏi:

1. Bạn thấy sản phẩm ở cửa hàng như thế nào?
   - Đa dạng, dễ chọn
   - Đầy đủ
   - Hạn chế
2. Giá cả sản phẩm:
   - Hợp lý
   - Cao
   - Thấp
3. Nhân viên tư vấn bán hàng:
   - Nhiệt tình, dễ chịu
   - Bình thường
   - Không thân thiện
4. Không gian, bố trí cửa hàng:
   - Gọn gàng, dễ tìm
   - Hơi lộn xộn
   - Khó tìm hàng
5. Bạn có quay lại lần sau không?
   - Có

- Không chắc
- Không

4. Mẫu khảo sát hài lòng tại cơ sở y tế (phòng khám, bệnh viện)

Tiêu đề:

> **Khảo sát sự hài lòng bệnh nhân – [Tên cơ sở y tế]**

Câu hỏi:

1. Bạn có hài lòng với thời gian chờ khám không?
   - Rất hài lòng
   - Tạm chấp nhận
   - Quá lâu
2. Bác sĩ/nhân viên y tế có tư vấn rõ ràng và tận tâm không?
   - Có
   - Bình thường
   - Không
3. Cơ sở vật chất:
   - Sạch sẽ, đầy đủ
   - Tạm ổn
   - Không tốt
4. Thái độ nhân viên tiếp đón:
   - Lịch sự, thân thiện
   - Bình thường
   - Thiếu chuyên nghiệp
5. Bạn có muốn giới thiệu cơ sở cho người khác không?
   - Có
   - Không chắc
   - Không
6. Góp ý hoặc phản ánh (nếu có):
   ➤ .....

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Thống Kê Mô Tả (Descriptive Statistics)

Phần lớn thông tin thống kê trên báo chí, tạp chí, báo cáo doanh nghiệp và các ấn phẩm khác đều là dữ liệu đã được **tổng hợp và trình bày** dưới dạng dễ hiểu cho người đọc. Những bản tóm tắt này — có thể ở dạng **bảng biểu, đồ thị hoặc số liệu** — được gọi là **thống kê mô tả (descriptive statistics)**.

Ví dụ minh họa tại thị trường Việt Nam

Xét lại bảng dữ liệu bạn cung cấp về **5 cổ phiếu niêm yết trên sàn chứng khoán Việt Nam** (VinGroup, Vietcombank, Hoa Phat, Masan Group, BIDV). Chúng ta có thể sử dụng **thống kê mô tả** để tóm tắt thông tin như sau:

1. Tóm tắt dạng bảng (Tabular Summary)

Giả sử chúng ta muốn xem **khối lượng giao dịch (Volume Traded)** của các cổ phiếu này:

Công ty	Khối lượng giao dịch (cổ phiếu)	Tỷ trọng (%)
VinGroup (VIC)	2,500,000	15.6%
Vietcombank (VCB)	3,500,000	21.9%
Hoa Phat (HPG)	4,200,000	26.3%
Masan Group (MSN)	3,000,000	18.8%
BIDV (BID)	2,800,000	17.5%
Tổng	**16,000,000**	**100%**

→ Từ bảng này, ta thấy **Hoa Phat (HPG)** có khối lượng giao dịch lớn nhất (26.3%), trong khi **VinGroup (VIC)** thấp nhất (15.6%).

2. Tóm tắt dạng đồ thị (Graphical Summary)

Một cách trực quan hơn, ta có thể dùng **biểu đồ cột (bar chart)** để so sánh khối lượng giao dịch:

VIC | ▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓ (2.5)

Biểu đồ khối lượng giao dịch 5 cổ phiếu (đơn vị: triệu cổ phiếu)

3. Tóm tắt bằng số liệu (Numerical Summary)

- **Trung bình (Mean)**:
  $\frac{2.5+3.5+4.2+3.0+2.8}{5} = 3.2$ triệu cổ phiếu.
- **Trung vị (Median)**: 3.0 triệu cổ phiếu (giá trị ở giữa khi sắp xếp).
- **Độ lệch chuẩn (Standard Deviation)**: ~0.7 triệu (đo lường độ phân tán).

Nhận xét

- **Hoa Phat (HPG)** chiếm **26.3%** tổng khối lượng giao dịch, cao nhất trong nhóm.
- **3 cổ phiếu (HPG, VCB, MSN)** chiếm **66.4%** tổng khối lượng → Thị trường có xu hướng tập trung vào một số mã nhất định.
- **VinGroup (VIC)** có khối lượng thấp nhất, có thể do giá cổ phiếu cao (100,000 VND/cổ phiếu).

Ứng dụng thực tế tại Việt Nam

- **Nhà đầu tư**: Dùng thống kê mô tả để so sánh thanh khoản giữa các cổ phiếu.
- **Công ty niêm yết**: Phân tích biến động giá và khối lượng giao dịch để đánh giá sự quan tâm của thị trường.
- **Cơ quan quản lý (UBCKNN, HOSE, HNX)**: Theo dõi xu hướng giao dịch để phát hiện bất thường.

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Suy luận thống kê (Statistical inference)

Trong nhiều trường hợp, chúng ta cần thu thập dữ liệu từ một nhóm lớn các đối tượng (cá nhân, doanh nghiệp, cử tri, hộ gia đình, sản phẩm, khách hàng, v.v.). Tuy nhiên, do hạn chế về thời gian, chi phí và các yếu tố khác, dữ liệu thường chỉ được thu thập từ một phần nhỏ của nhóm này.

- **Tổng thể (Population)**: Toàn bộ nhóm đối tượng cần nghiên cứu.
- **Mẫu (Sample)**: Một tập hợp con được chọn từ tổng thể.

Quá trình thu thập dữ liệu từ toàn bộ tổng thể được gọi là **điều tra toàn bộ (census)**, trong khi việc thu thập từ một mẫu được gọi là **điều tra mẫu (sample survey)**. Một trong những đóng góp quan trọng của thống kê là sử dụng dữ liệu mẫu để **ước lượng** và **kiểm định giả thuyết** về đặc điểm của tổng thể, thông qua quá trình gọi là **suy luận thống kê (statistical inference)**.

Ví dụ minh họa

Giả sử **Công ty Bóng đèn Điện Quang** muốn cải tiến tuổi thọ của bóng đèn LED dân dụng. Nhóm nghiên cứu phát triển một loại chip LED mới, và tổng thể ở đây là **tất cả bóng đèn có thể được sản xuất với công nghệ mới này**.

Để đánh giá hiệu quả, công ty sản xuất thử nghiệm **500 bóng đèn** (mẫu) và ghi lại số giờ hoạt động cho đến khi hỏng. Kết quả thu được như sau:

Dữ liệu mẫu (500 bóng đèn)

- Tuổi thọ trung bình: **25,000 giờ**
- Độ lệch chuẩn: **2,000 giờ**

Ước lượng thống kê

1. **Ước lượng điểm (Point Estimate)**:
   - Dựa trên mẫu, tuổi thọ trung bình của bóng đèn mới là **25,000 giờ**.
   - Đây là ước lượng cho tuổi thọ trung bình của **toàn bộ sản phẩm** nếu sản xuất đại trà.
2. **Khoảng tin cậy (Interval Estimate)**:
   - Với **độ tin cậy 95%**, sai số ước lượng là **±500 giờ**.
     → Khoảng ước lượng: **24,500 – 25,500 giờ**.
   - Nghĩa là, có 95% khả năng tuổi thọ thực tế của toàn bộ đèn LED mới nằm trong khoảng này.

Quy trình suy luận thống kê

1. **Xác định tổng thể**: Tất cả bóng đèn LED sản xuất bằng công nghệ mới.
2. **Thu thập mẫu**: 500 bóng đèn thử nghiệm.
3. **Tính toán thống kê mẫu**: Trung bình, độ lệch chuẩn.
4. **Suy luận về tổng thể**:
   - Ước lượng điểm: 25,000 giờ.
   - Khoảng tin cậy: 24,500 – 25,500 giờ.

Ứng dụng thực tế tại Việt Nam

1. **Kiểm tra chất lượng sản phẩm**:
   - Các công ty như **VinFast, TH True Milk, Hòa Phát** thường dùng suy luận thống kê để đánh giá độ bền, độ an toàn trước khi sản xuất hàng loạt.
2. **Nghiên cứu thị trường**:
   - Ví dụ: Công ty **The Coffee House** muốn khảo sát mức độ hài lòng của khách hàng. Thay vì hỏi tất cả, họ chỉ khảo sát **1,000 khách hàng** và suy luận cho toàn bộ thị trường.
3. **Dự báo kinh tế**:
   - Tổng cục Thống kê Việt Nam (GSO) thường dùng **điều tra mẫu** để ước lượng GDP, tỷ lệ thất nghiệp, lạm phát.

Tóm tắt

- **Tổng thể (Population)**: Nhóm đối tượng cần nghiên cứu.
- **Mẫu (Sample)**: Tập hợp con đại diện, giúp tiết kiệm chi phí.
- **Suy luận thống kê**: Dùng dữ liệu mẫu để ước lượng hoặc kiểm định giả thuyết về tổng thể.
- **Độ chính xác**: Luôn kèm theo **sai số (margin of error)** và **độ tin cậy (confidence level)**.

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Ứng dụng công nghệ trong phân tích dữ liệu tại Việt Nam

1. Vai trò của máy tính trong phân tích thống kê

Phân tích thống kê thường xử lý khối lượng dữ liệu lớn, do đó các nhà phân tích thường sử dụng **phần mềm chuyên dụng** để:

- **Tự động hóa tính toán** (trung bình, độ lệch chuẩn, hồi quy...)
- **Xử lý dữ liệu nhanh chóng** so với phương pháp thủ công
- **Trực quan hóa dữ liệu** bằng biểu đồ, báo cáo động

Ví dụ: Tính tuổi thọ trung bình của 500 bóng đèn Điện Quang (như ví dụ trước) sẽ mất hàng giờ nếu tính tay, nhưng chỉ cần **vài giây** với phần mềm.

2. Các công cụ phổ biến tại Việt Nam

a. Phần mềm quốc tế

- **Excel**: Được dùng rộng rãi nhờ giao diện thân thiện, tích hợp sẵn hàm thống kê (AVERAGE, STDEV, CORREL...).
- **SPSS**: Phổ biến trong nghiên cứu xã hội học, y tế (khảo sát ý kiến, phân tích ANOVA).
- **R & Python**: Miễn phí, mạnh về xử lý dữ liệu lớn và AI, được các công ty công nghệ như VinBigData, FPT ứng dụng.

b. Giải pháp Việt Nam

- **Phần mềm STATA bản địa hóa**: Một số trường ĐH (KTQD, BKHN) phát triển giao diện tiếng Việt để giảng dạy.
- **Nền tảng điện toán đám mây**: Như **VNG Cloud**, **Viettel AI** hỗ trợ xử lý dữ liệu doanh nghiệp.

3. Xu hướng hiện nay

- **Tự động hóa báo cáo**: Các ngân hàng (Vietcombank, Techcombank) dùng **Power BI** để phân tích giao dịch theo thời gian thực.

- **AI trong dự báo**: Tập đoàn Vingroup ứng dụng **machine learning** để dự đoán nhu cầu thị trường ô tô.
- **Mở rộng đào tạo**: Các khóa học online về Data Science (trên Coursera, Funix) thu hút hàng nghìn người học tại Việt Nam.

4. Hướng dẫn thực hành

Để phân tích dữ liệu như ví dụ bóng đèn Điện Quang:

1. **Nhập liệu**: Lưu file Excel (định dạng .xlsx) hoặc CSV.
2. **Phần mềm khuyến nghị**:
   - Người mới: **Excel** (dùng Data Analysis ToolPak).
   - Nâng cao: **R** (code mẫu: `mean(data$lifetime)`).
3. **Tài nguyên Việt Nam**:
   - Kho dữ liệu mẫu từ **Tổng cục Thống kê** (gso.gov.vn).
   - Diễn đàn **R Vietnam** trên Facebook để trao đổi chuyên môn.

Kết luận

Việc ứng dụng máy tính và phần mềm đã cách mạng hóa phân tích thống kê tại Việt Nam, giúp:
□□ **Tiết kiệm thời gian**
□□ **Nâng cao độ chính xác**
□□ **Mở ra cơ hội** trong thời đại 4.0

Gợi ý: Các doanh nghiệp vừa và nhỏ (SMEs) có thể bắt đầu với Excel hoặc Google Sheets trước khi chuyển sang công cụ phức tạp hơn.

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Khai phá dữ liệu

Giới thiệu về Khai phá dữ liệu

Với sự hỗ trợ của **máy đọc thẻ từ, máy quét mã vạch, hệ thống POS (điểm bán hàng)**, các doanh nghiệp ngày nay thu thập một lượng dữ liệu khổng lồ mỗi ngày. Ngay cả một **quán cà phê nhỏ** sử dụng phần mềm order cũng có thể tích lũy dữ liệu đáng kể về thói quen khách hàng.

- **Tại Việt Nam**, các tập đoàn như **VinCommerce (VinMart), Thegioididong, Shopee** ghi nhận hàng triệu giao dịch mỗi ngày.
- **Ví dụ**:
  - **Momo** xử lý ~10 triệu giao dịch/ngày (2023).
  - **Shopee Vietnam** ghi nhận hơn 2 triệu đơn hàng/ngày trong các đợt sale.

Kho dữ liệu (Data Warehousing)

- **Định nghĩa**: Quá trình **thu thập, lưu trữ và quản lý** dữ liệu quy mô lớn.
- **Ứng dụng tại Việt Nam**:
  - **Ngân hàng (Vietcombank, Techcombank)**: Lưu trữ dữ liệu giao dịch, lịch sử tín dụng.
  - **Bán lẻ (VinMart, Bach Hóa Xanh)**: Theo dõi hành vi mua sắm qua hệ thống POS.

Khai phá dữ liệu là gì?

Là quá trình **phân tích dữ liệu** để phát hiện xu hướng, mẫu hình ẩn, hỗ trợ ra quyết định kinh doanh.

Công nghệ sử dụng

- **Thống kê (Statistics)**: Phân hồi quy, phân cụm.
- **Trí tuệ nhân tạo (AI)**: Học máy (Machine Learning), cây quyết định.
- **Ví dụ tại Việt Nam**:
  - **Tiki** dùng **recommendation engine** đề xuất sản phẩm dựa trên lịch sử mua hàng.
  - **VinID** phân tích dữ liệu tiêu dùng để gửi voucher cá nhân hóa.

Ứng dụng thực tế tại Việt Nam

1. Bán lẻ & Thương mại điện tử

- **Shopee/Lazada**:
  - Phân tích **"Frequently Bought Together"** (ví dụ: Khách mua điện thoại thường mua thêm ốp lưng).

    - Tối ưu **flash sale** dựa trên dữ liệu mua hàng đỉnh điểm.
- **VinMart**:
    - Dự báo nhu cầu sản phẩm theo mùa (ví dụ: tăng nhập bia vào mùa hè).

2. Ngân hàng & Tài chính

- **Fraud Detection (Phát hiện gian lận)**:
    - **VPBank** sử dụng AI để nhận diện giao dịch thẻ tín dụng bất thường.
- **Scoring tín dụng**:
    - **FE Credit** phân tích hành vi tiêu dùng để đánh giá rủi ro cho vay.

3. Viễn thông (Viettel, Vinaphone)

- **Phân tích cuộc gọi**: Phát hiện nhóm khách hàng có nguy cơ chuyển mạng (churn prediction).
- **Tối ưu gói cước**: Đề xuất gói data phù hợp với từng nhóm người dùng.

Thách thức & Giải pháp

1. Độ tin cậy mô hình (Model Reliability)

- **Vấn đề**: Mô hình chạy tốt trên dữ liệu mẫu nhưng có thể sai lệch khi áp dụng thực tế.
- **Giải pháp**:
    - Chia dữ liệu thành **train set** (70%) và **test set** (30%).
    - **Ví dụ**: Các ngân hàng kiểm tra mô hình dự đoán rủi ro trước khi triển khai.

2. Hiểu sai quan hệ nhân quả (Overfitting)

- **Ví dụ**: Nếu phân tích dữ liệu thời tiết và doanh số kem, có thể kết luận **"mưa nhiều làm tăng bán kem"** (sai logic).
- **Giải pháp**: Kết hợp **kiểm định thống kê** và hiểu biết chuyên ngành.

Xu hướng tại Việt Nam

1. **AI & Big Data**:
    - Các startup như **Trusting Social, VHTSoft** (phân tích tín dụng) sử dụng data mining để đánh giá rủi ro.
2. **Personalized Marketing**:
    - **The Coffee House** dùng dữ liệu member để gửi voucher cá nhân hóa.
3. **Chính phủ số**:
    - **Cổng Dịch vụ công Quốc gia** phân tích dữ liệu để tối ưu thủ tục hành chính.

Kết luận

Khai phá dữ liệu đang trở thành **công cụ chiến lược** tại Việt Nam, giúp doanh nghiệp:
Tăng doanh thu (qua recommendation systems)
Giảm rủi ro (phát hiện gian lận)
Tối ưu vận hành (dự báo nhu cầu)

Tuy nhiên, cần **kết hợp thống kê truyền thống và AI** để tránh sai lệch trong phân tích!

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Tóm tắt

Khái niệm cơ bản

Thống kê là **nghệ thuật và khoa học** thu thập, phân tích, trình bày và diễn giải dữ liệu. Đây là môn học bắt buộc với sinh viên các ngành kinh tế và quản trị kinh doanh. Chương này đã giới thiệu các ứng dụng thống kê điển hình trong lĩnh vực kinh doanh.

Dữ liệu và thang đo

- **Dữ liệu (Data)**: Là các con số và thông tin được thu thập để phân tích.
- **Quan sát (Observation)**: Tập hợp các giá trị đo lường cho một đơn vị nghiên cứu cụ thể.

4 thang đo trong thống kê

1. **Định danh (Nominal)**: Dùng nhãn hoặc tên để phân loại (ví dụ: ngành nghề, giới tính).
2. **Thứ bậc (Ordinal)**: Có thứ tự nhưng khoảng cách không đều (ví dụ: xếp hạng khách hàng VIP, Gold, Silver).
3. **Khoảng cách (Interval)**: Có thứ tự và khoảng cách đều, không có điểm 0 tuyệt đối (ví dụ: nhiệt độ °C).
4. **Tỷ lệ (Ratio)**: Có đầy đủ tính chất của interval và có điểm 0 tuyệt đối (ví dụ: doanh thu, tuổi tác).

Phân loại dữ liệu

- **Dữ liệu định tính (Categorical data)**:
  - Sử dụng thang đo nominal hoặc ordinal
  - Ví dụ: Loại hình doanh nghiệp (TNHH, Cổ phần), Xếp hạng tín nhiệm
- **Dữ liệu định lượng (Quantitative data)**:
  - Sử dụng thang đo interval hoặc ratio
  - Ví dụ: Lợi nhuận quý (tỷ đồng), Số nhân viên
  - Có thể áp dụng các phép toán số học

Phương pháp thống kê

1. **Thống kê mô tả (Descriptive statistics)**:
   - Tổng hợp dữ liệu qua bảng biểu, đồ thị hoặc số liệu
   - Ví dụ: Báo cáo doanh thu theo quý của Vinamilk
2. **Suy luận thống kê (Statistical inference)**:
   - Sử dụng dữ liệu mẫu để ước lượng hoặc kiểm định cho tổng thể
   - Ví dụ: Khảo sát 1,000 hộ gia đình để dự báo chi tiêu Tết 2024

Công nghệ hỗ trợ

- **Phần mềm thống kê**: SPSS, R, Python giúp xử lý dữ liệu phức tạp
- **Khai phá dữ liệu (Data mining)**: Ứng dụng AI để phát hiện xu hướng ẩn trong big data

Thuật ngữ chính

Tiếng Anh	Tiếng Việt	Ví dụ ứng dụng tại VN
Population	Tổng thể	Toàn bộ DN niêm yết HOSE
Sample	Mẫu	30 công ty vốn hóa lớn nhất
Time series data	Dữ liệu chuỗi thời gian	Biến động giá cổ phiếu VIC 5 năm
Cross-sectional data	Dữ liệu chéo	Khảo sát thu nhập hộ gia đình 2023
Data mining	Khai phá dữ liệu	Phân tích hành vi mua hàng trên Shopee

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Bài Tập: Sự Khác Biệt Giữa Thống Kê Là Con Số và Thống Kê Là Một Ngành Học

1. Thống kê là các con số (Statistics as Numerical Facts)

- **Định nghĩa**: Là những **dữ liệu số** cụ thể được thu thập để mô tả sự vật, hiện tượng.
- **Vai trò**:
  - Cung cấp thông tin định lượng về một vấn đề.
  - Thường xuất hiện trong báo cáo, tin tức, nghiên cứu thị trường.
- **Ví dụ**:
  - "GDP Việt Nam quý I/2024 tăng **5,66%** so với cùng kỳ năm trước" *(Tổng cục Thống kê)*.
  - "Doanh thu Shopee Việt Nam đạt **1,2 tỷ USD** năm 2023".

2. Thống kê là một ngành học (Statistics as a Discipline)

- **Định nghĩa**: Là **môn khoa học** bao gồm các phương pháp thu thập, phân tích, diễn giải và trình bày dữ liệu.
- **Vai trò**:
  - Cung cấp công cụ để rút ra kết luận từ dữ liệu.
  - Ứng dụng trong kinh tế, y tế, khoa học xã hội, AI...
- **Ví dụ** :
  - **Phân tích hồi quy** để dự báo lạm phát (Ngân hàng Nhà nước).
  - **Kiểm định A/B testing** tối ưu giao diện app Momo.

Bảng So Sánh Chi Tiết

Tiêu chí	Thống kê là con số	Thống kê là ngành học
Bản chất	Dữ liệu đầu ra (kết quả)	Quy trình phân tích để tạo ra kết quả
Mục đích	Mô tả hiện trạng	Suy luận, dự báo, ra quyết định

Tiêu chí	Thống kê là con số	Thống kê là ngành học
Ví dụ thực tế tại VN	"70% người dùng Internet mua sắm online"	Sử dụng **phân cụm (clustering)** để phân khúc khách hàng của Tiki
Tính ứng dụng	Truyền thông, báo cáo	Nghiên cứu thị trường, phát triển sản phẩm

Kết Luận

- **Thống kê là con số**: Cung cấp **"câu trả lời"** cụ thể nhưng không giải thích "tại sao".
- **Thống kê là ngành học**: Cung cấp **"công cụ"** để tìm câu trả lời và đưa ra quyết định khoa học.

Ứng dụng trong kinh doanh :

- Các con số (ví dụ: "Doanh thu VinFast tăng 20%") giúp đánh giá hiệu quả.
- Phương pháp thống kê (ví dụ: **dự báo chuỗi thời gian**) giúp VinFast lập kế hoạch sản xuất.

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Bài tập: Thống kê về Khách sạn tại Việt Nam

Phân tích **Top 10 khách sạn sang trọng nhất Việt Nam năm 2023** theo tạp chí Du lịch Heritage:

Tên khách sạn	Địa điểm	Hạng sao	Giá phòng (USD/đêm)	Loại hình
InterContinental Danang Sun Peninsula Resort	Đà Nẵng	5	450	Resort biển
The Reverie Saigon	TP.HCM	5	400	Khách sạn thành phố
JW Marriott Hanoi	Hà Nội	5	350	Khách sạn thành phố
Amanoi Resort	Ninh Thuận	5	600	Resort nghỉ dưỡng
Six Senses Ninh Van Bay	Khánh Hòa	5	550	Resort biệt lập
Four Seasons Resort The Nam Hai	Quảng Nam	5	500	Resort biển
Sofitel Legend Metropole Hanoi	Hà Nội	5	300	Khách sạn di sản
Banyan Tree Lang Co	Thừa Thiên Huế	5	480	Resort golf
Park Hyatt Saigon	TP.HCM	5	380	Khách sạn boutique
Anantara Mui Ne Resort	Bình Thuận	5	420	Resort biển

a. Số lượng phần tử (elements) trong tập dữ liệu

b. Số lượng biến số (variables) trong tập dữ liệu

c. Phân loại biến định tính và định lượng

d. Thang đo cho từng biến số

e. Tính số phòng trung bình của 10 khách sạn này?

f. Nếu tỷ giá hối đoái:
1 EUR = 1.3149 USD

1 EUR = 0.8986 GBP
Hãy tính giá phòng trung bình bằng EUR.

g. Tính tỷ lệ phần trăm khách sạn tọa lạc tại Đà Nẵng?
(Gợi ý: Đếm số khách sạn ở Đà Nẵng chia tổng số khách sạn)

h. Tính tỷ lệ phần trăm khách sạn có 100 phòng hoặc ít hơn?
(Gợi ý: Đếm số khách sạn có số phòng ≤ 100)

Giải bài tập

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Bài tập Thống kê về Hệ thống Âm thanh Thông minh

Dữ liệu về hệ thống âm thanh phổ biến

Dưới đây là thông tin **5 hệ thống âm thanh thông minh bán chạy nhất** tại Việt Nam năm 2024:

Bài tập 1

Tên sản phẩm	Đánh giá (5★)	Giá (triệu VND)	Bluetooth	WiFi	Công suất (W)	Trợ lý ảo
Sony HT-A5000	4.8	15.9	Y	Y	500	Google Assistant
LG S95QR	4.7	22.5	Y	Y	610	Alexa
Samsung HW-Q990B	4.9	18.7	Y	Y	656	Bixby
JBL Bar 1000	4.6	12.3	Y	N	880	Không
Bose Smart Soundbar 900	4.5	16.8	Y	Y	450	Alexa

> *Ghi chú:
>
> - "Y" = Có tính năng này
> - "N" = Không có tính năng này*

Câu hỏi bài tập

1. **Có bao nhiêu phần tử (elements) trong tập dữ liệu này?**
   (Gợi ý: Đếm số dòng dữ liệu)
2. **Tổng thể (population) trong nghiên cứu này là gì?**
   (Gợi ý: Xác định nhóm đối tượng lớn hơn mà mẫu này đại diện)
3. **Tính công suất đầu ra trung bình của mẫu hệ thống âm thanh này.**
   (Gợi ý: Tính tổng công suất chia cho số lượng sản phẩm)
4. **Tính tỷ lệ phần trăm hệ thống có hỗ trợ trợ lý ảo.**
   (Gợi ý: Đếm số sản phẩm có trợ lý ảo chia tổng số sản phẩm)

5. **Có bao nhiêu hệ thống có giá dưới 18 triệu VND?**
   (Gợi ý: Đếm số sản phẩm thỏa điều kiện)

> *Lưu ý: Bài tập này giúp rèn luyện kỹ năng phân tích dữ liệu thực tế trong lĩnh vực công nghệ - một ngành đang phát triển mạnh tại Việt Nam.*

Bài tập 2

Tên sản phẩm	Đánh giá (5★)	Giá (triệu VND)	Bluetooth	MP3 Player	Công suất (W)	Loại
Sony HT-A5000	4.8	15.9	Y	Y	500	Soundbar
LG S95QR	4.7	22.5	Y	N	610	Home Theater
Samsung HW-Q990B	4.9	18.7	Y	Y	656	Soundbar
JBL Bar 1000	4.6	12.3	Y	N	880	Soundbar
Bose Smart Soundbar 900	4.5	16.8	Y	Y	450	Soundbar
Harman Kardon Citation	4.3	14.2	Y	N	400	Bookshelf
Sonos Arc	4.8	21.0	Y	Y	500	Soundbar
Philips TAB7005	4.2	10.9	Y	N	300	Soundbar

Câu hỏi bài tập

1. **Có bao nhiêu biến số (variables) trong tập dữ liệu này?**
2. **Biến số nào là định lượng (quantitative) và biến số nào là định tính (categorical)?**
3. **Tính tỷ lệ phần trăm hệ thống âm thanh có đánh giá từ 4 sao trở lên?**
4. **Tính tỷ lệ phần trăm hệ thống âm thanh có tích hợp MP3 Player?**

Bài tập: Thống kê Thông tin các công ty niêm yết trên HOSE (2024)

Dưới đây là bảng dữ liệu mẫu phù hợp với thị trường Việt Nam 2024 để áp dụng cho bài tập phân loại biến số:

Bảng : Thông tin các công ty niêm yết trên HOSE (2024)

Mã CK	Tên công ty	Ngành nghề	Vốn hóa (tỷ VNĐ)	EPS (VNĐ)	Tăng trưởng DT (%)	Xếp hạng tín nhiệm
VIC	Vingroup	Bất động sản	250,000	2,500	12.5	AAA
VCB	Vietcombank	Ngân hàng	420,000	4,500	8.2	AA+
HPG	Hòa Phát	Thép	300,000	3,800	15.7	AA
MSN	Masan Group	Tiêu dùng nhanh	120,000	1,200	10.3	A+
FPT	FPT Corporation	Công nghệ thông tin	180,000	5,600	22.8	AAA
MWG	Thế Giới Di Động	Bán lẻ điện tử	95,000	3,200	7.5	A

1. Phân loại các biến số (định tính/định lượng)
2. Xác định thang đo tương ứng
3. Thực hành tính toán các chỉ số thống kê cơ bản
4. Phân tích mối quan hệ giữa các biến

1. Phân loại biến số và xác định thang đo

Biến số	Loại biến	Thang đo	Giải thích
Mã CK	Định tính	Danh nghĩa (Nominal)	Dùng để nhận diện, không có thứ tự ưu tiên (VIC, VCB...)
Tên công ty	Định tính	Danh nghĩa (Nominal)	Tên gọi phân loại, không thể xếp hạng
Ngành nghề	Định tính	Danh nghĩa (Nominal)	Phân loại ngành (BĐS, ngân hàng...), không có thứ bậc
Vốn hóa (tỷ VNĐ)	Định lượng	Tỷ lệ (Ratio)	Có điểm 0 tuyệt đối, có thể tính tỷ lệ (ví dụ: 420,000 gấp đôi 210,000)
EPS (VNĐ)	Định lượng	Tỷ lệ (Ratio)	Có điểm 0 tuyệt đối, phép chia có ý nghĩa (ví dụ: EPS 5,600 gấp 2 lần 2,800)
Tăng trưởng DT (%)	Định lượng	Khoảng cách (Interval)	Không có điểm 0 tuyệt đối (tăng trưởng âm/vô nghĩa), nhưng khoảng cách đều
Xếp hạng tín nhiệm	Định tính	Thứ bậc (Ordinal)	Có thứ tự (AAA > AA+ > AA > A+ > A) nhưng khoảng cách không đồng nhất

---

2. Tính toán các chỉ số thống kê cơ bản

a. Vốn hóa thị trường (tỷ VNĐ):

- **Trung bình (Mean):**
  $\frac{250+420+300+120+180+95}{6} = \frac{1,365}{6} \approx 227.5$ tỷ VNĐ
- **Trung vị (Median):**
  Sắp xếp: 95, 120, 180, 250, 300, 420 → Trung vị = $\frac{180+250}{2} = 215$ tỷ VNĐ
- **Độ lệch chuẩn (SD):** ≈ 120.8 tỷ VNĐ (đo độ phân tán)

b. EPS (VNĐ):

- **Trung bình:** $\frac{2,500+4,500+3,800+1,200+5,600+3,200}{6} \approx 3,467$ VNĐ
- **Min-Max:** 1,200 (Masan) – 5,600 (FPT)

c. Tăng trưởng doanh thu (%):

- **Mốt (Mode):** Không có giá trị lặp lại → Không có mốt.
- **Phân phối:**
  - Công ty công nghệ (FPT) có tăng trưởng cao nhất (22.8%).
  - Ngân hàng (VCB) và bán lẻ (MWG) tăng trưởng thấp hơn (<10%).

3. Phân tích mối quan hệ giữa các biến

a. Vốn hóa vs. EPS:

- **Nhận xét:**
  - Công ty vốn hóa lớn (VCB: 420,000 tỷ) không nhất thiết có EPS cao nhất (FPT: 5,600 VNĐ dù vốn hóa chỉ 180,000 tỷ).
  - **Nguyên nhân:** EPS phụ thuộc vào lợi nhuận và số cổ phiếu lưu hành.

b. Ngành nghề vs. Tăng trưởng doanh thu:

- **Xu hướng:**
  - **Công nghệ (FPT):** Tăng trưởng cao nhất (22.8%) do nhu cầu chuyển đổi số.
  - **Bán lẻ (MWG):** Tăng trưởng thấp (7.5%) do cạnh tranh khốc liệt.

c. Xếp hạng tín nhiệm vs. Vốn hóa:

- **Quy luật:**
  - Công ty vốn hóa lớn (VIC, VCB) thường có xếp hạng tín nhiệm cao (AAA, AA+).
  - **Ngoại lệ:** FPT (vốn hóa trung bình nhưng xếp hạng AAA) nhờ uy tín ngành công nghệ.

4. Bài tập thực hành thêm

Câu 1: Tính tỷ lệ công ty có vốn hóa trên 200,000 tỷ VNĐ.
Gợi ý: 4/6 ≈ 66.67% (VIC, VCB, HPG, FPT).

Câu 2: So sánh EPS trung bình giữa ngành ngân hàng (VCB) và công nghệ (FPT).
Gợi ý: Ngân hàng (4,500) < Công nghệ (5,600).

Câu 3: Vẽ biểu đồ scatter plot giữa Vốn hóa và Tăng trưởng doanh thu. Nhận xét?
Gợi ý: Không có tương quan rõ ràng, ví dụ FPT vốn hóa trung bình nhưng tăng trưởng cao nhất.

Kết luận

- **Biến định lượng (Vốn hóa, EPS)** phù hợp để tính toán chỉ số trung bình, độ phân tán.
- **Biến định tính (Ngành nghề, Xếp hạng)** dùng để phân nhóm và so sánh đặc điểm.

- **Ứng dụng:** Nhà đầu tư có thể kết hợp phân tích đa biến để chọn cổ phiếu tiềm năng (ví dụ: FPT dù vốn hóa không lớn nhưng có EPS cao và tăng trưởng mạnh).

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Bài tập: phân loại (categorical) hay định lượng (quantitative) và chỉ ra thang đo (measurement scale)

Bảng dữ liệu mẫu (dựa trên khảo sát 10 cá nhân):

ID	Thu nhập hàng năm (triệu VND)	Trình độ học vấn	Loại phương tiện di chuyển	Số giờ làm việc/tuần	Tình trạng hôn nhân
1	240	Cử nhân	Xe máy	40	Đã kết hôn
2	180	THPT	Xe đạp	35	Độc thân
3	300	Thạc sĩ	Ô tô	45	Đã kết hôn
4	150	THPT	Đi bộ	30	Độc thân
5	200	Cử nhân	Xe máy	42	Ly hôn
6	270	Thạc sĩ	Ô tô	50	Đã kết hôn
7	120	THPT	Xe buýt	25	Độc thân
8	350	Tiến sĩ	Ô tô	48	Đã kết hôn
9	190	Cử nhân	Xe máy	38	Độc thân
10	220	Cử nhân	Xe buýt	40	Ly hôn

Yêu cầu

Hãy xác định mỗi biến sau đây là biến **phân loại (categorical)** hay **định lượng (quantitative)** và chỉ ra **thang đo (measurement scale)** của nó.

a. Thu nhập hàng năm (Annual income).
b. Trình độ học vấn (Education level).

c. Loại phương tiện di chuyển (Type of transportation).
d. Số giờ làm việc mỗi tuần (Number of working hours per week).

Đáp án gợi ý:

a. **Thu nhập hàng năm (Annual income)**:

- Loại biến: Định lượng (Quantitative).
- Thang đo: Tỷ lệ (Ratio scale) – có điểm 0 tuyệt đối, có thể so sánh tỷ lệ (ví dụ: 300 triệu gấp 2 lần 150 triệu).

b. **Trình độ học vấn (Education level)**:

- Loại biến: Phân loại (Categorical).
- Thang đo: Thứ tự (Ordinal scale) – các giá trị có thứ tự (THPT < Cử nhân < Thạc sĩ < Tiến sĩ), nhưng khoảng cách giữa các mức không đồng đều.

c. **Loại phương tiện di chuyển (Type of transportation)**:

- Loại biến: Phân loại (Categorical).
- Thang đo: Danh nghĩa (Nominal scale) – các giá trị chỉ phân loại, không có thứ tự (xe máy, ô tô, xe đạp, v.v.).

d. **Số giờ làm việc mỗi tuần (Number of working hours per week)**:

- Loại biến: Định lượng (Quantitative).
- Thang đo: Tỷ lệ (Ratio scale) – có điểm 0 tuyệt đối, có thể thực hiện các phép toán như cộng, trừ, nhân, chia.

e. **Tình trạng hôn nhân (Marital status)**:

- Loại biến: Phân loại (Categorical).
- Thang đo: Danh nghĩa (Nominal scale) – các giá trị như độc thân, đã kết hôn, ly hôn chỉ phân loại, không có thứ tự.

Bài tập thống kê: Phân tích thu nhập ròng của Volkswagen (2016–2024)

Bảng dữ liệu: Thu nhập ròng của Volkswagen (tỷ USD)

Năm	Thu nhập ròng (tỷ USD)
2016	5.71
2017	12.92
2018	14.32
2019	15.54
2020	9.61
2021	17.56
2022	15.66
2023	17.33
2024	11.60

Ghi chú: Dữ liệu thu nhập ròng được lấy từ MacroTrends và báo cáo tài chính của Volkswagen, làm tròn đến hai chữ số thập phân.

Câu hỏi

a. Dữ liệu này là **phân loại (categorical)** hay **định lượng (quantitative)**?
b. Dữ liệu này là **chuỗi thời gian (time series)** hay **dữ liệu chéo (cross-sectional)**?
c. Biến quan tâm (variable of interest) là gì?
d. Nhận xét về xu hướng thu nhập ròng của Volkswagen qua các năm. Bạn dự đoán thu nhập ròng

sẽ **tăng** hay **giảm** vào năm 2025?

Đáp án gợi ý

a. **Dữ liệu là định lượng (quantitative)**:
Thu nhập ròng được đo bằng số tiền (tỷ USD), là một giá trị số có thể thực hiện các phép toán như cộng, trừ, nhân, chia. Do đó, đây là dữ liệu định lượng.

b. **Dữ liệu là chuỗi thời gian (time series)**:
Dữ liệu được thu thập theo thời gian (từ năm 2016 đến 2024) cho cùng một đối tượng (Volkswagen). Chuỗi thời gian ghi lại sự thay đổi của một biến qua các mốc thời gian liên tiếp, khác với dữ liệu chéo(so sánh nhiều đối tượng tại một thời điểm).

c. **Biến quan tâm**:
Biến quan tâm là **thu nhập ròng (net income)** của Volkswagen, được đo bằng tỷ USD.

d. **Nhận xét về xu hướng và dự đoán cho năm 2025**:

- **Xu hướng**: Từ năm 2016 đến 2024, thu nhập ròng của Volkswagen có xu hướng biến động:
  - Tăng mạnh từ 5.71 tỷ USD (2016) lên 15.54 tỷ USD (2019), cho thấy giai đoạn tăng trưởng ổn định.
  - Giảm đáng kể xuống 9.61 tỷ USD vào năm 2020, có thể do tác động của đại dịch COVID-19 ảnh hưởng đến ngành ô tô.
  - Phục hồi mạnh mẽ vào năm 2021 (17.56 tỷ USD) và duy trì ở mức cao vào năm 2022 (15.66 tỷ USD) và 2023 (17.33 tỷ USD).
  - Tuy nhiên, năm 2024 ghi nhận sự sụt giảm đáng kể xuống 11.60 tỷ USD, giảm 33.06% so với năm 2023, có thể do các yếu tố như chi phí sản xuất tăng, cạnh tranh thị trường hoặc các vấn đề kinh tế toàn cầu.
- **Dự đoán cho năm 2025**: Dựa trên xu hướng giảm mạnh vào năm 2024 và các thách thức kinh tế toàn cầu (như nhu cầu giảm ở châu Âu hoặc chi phí nguyên liệu tăng), có khả năng thu nhập ròng của Volkswagen sẽ **tiếp tục giảm** hoặc duy trì ở mức thấp vào năm 2025, trừ khi có các yếu tố tích cực như cải thiện thị trường hoặc chiến lược kinh doanh mới. Tuy nhiên, dự đoán này mang tính suy đoán và cần thêm dữ liệu thực tế để xác nhận.

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Bài tập: Thống kê về du khách tại Việt Nam

Cục Du lịch Quốc gia Việt Nam thu thập dữ liệu về du khách đến Việt Nam. Dưới đây là các câu hỏi được đưa ra trong bảng câu hỏi phát cho hành khách trên các chuyến bay nội địa và quốc tế đến Việt Nam vào năm 2025.

1. Chuyến đi này đến Việt Nam là lần thứ: 1, 2, 3, 4, v.v.
2. Lý do chính cho chuyến đi này là: (10 danh mục bao gồm du lịch nghỉ dưỡng, hội nghị, tuần trăng mật, thăm thân).
3. Nơi tôi dự định lưu trú: (10 danh mục bao gồm khách sạn, căn hộ, nhà người thân, homestay).
4. Tổng số ngày lưu trú tại Việt Nam.

Câu hỏi:

a. Dân số được nghiên cứu là gì?
b. Việc sử dụng bảng câu hỏi có phải là cách tốt để tiếp cận dân số hành khách trên các chuyến bay đến Việt Nam không?
c. Nhận xét về từng câu hỏi trong bốn câu hỏi trên, liệu câu hỏi đó cung cấp dữ liệu định tính (categorical) hay định lượng (quantitative).

Đáp án gợi ý:

a. **Dân số được nghiên cứu**: Tất cả du khách đến Việt Nam bằng đường hàng không trong năm 2025.
b. **Đánh giá việc sử dụng bảng câu hỏi**:

- Ưu điểm: Bảng câu hỏi là cách hiệu quả để thu thập dữ liệu từ một lượng lớn hành khách trong thời gian ngắn, đặc biệt khi họ đang trên chuyến bay.
- Nhược điểm: Có thể bỏ sót một số hành khách không muốn trả lời hoặc không hiểu ngôn ngữ của bảng câu hỏi (ví dụ: du khách quốc tế). Ngoài ra, dữ liệu có thể không đại diện cho du khách đến bằng các phương tiện khác (tàu, đường bộ).
  c. **Nhận xét về từng câu hỏi**:
- Câu 1: **Định lượng** – Số lần đến Việt Nam là một biến số đếm (1, 2, 3, ...).
- Câu 2: **Định tính** – Lý do chuyến đi thuộc các danh mục cố định (du lịch, hội nghị, tuần trăng mật, ...).

- Câu 3: **Định tính** – Nơi lưu trú thuộc các danh mục cố định (khách sạn, homestay, ...).
- Câu 4: **Định lượng** – Tổng số ngày lưu trú là một biến số liên tục hoặc số đếm.

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Bài tập: Thống kê về quyết định tăng lương

Một quản lý của một tập đoàn lớn tại Việt Nam đề xuất tăng lương 25 triệu VND/tháng để giữ chân một nhân viên có giá trị, nhằm ngăn họ chuyển sang công ty khác. Những nguồn dữ liệu nội bộ và bên ngoài nào có thể được sử dụng để quyết định xem việc tăng lương này có phù hợp hay không?

Câu hỏi:

Hãy liệt kê và mô tả các nguồn dữ liệu nội bộ và bên ngoài có thể được sử dụng để đánh giá đề xuất tăng lương này.

Đáp án gợi ý:

1. Nguồn dữ liệu nội bộ:

- **Hồ sơ nhân sự**:
  Dữ liệu về hiệu suất làm việc của nhân viên (đánh giá KPI, đóng góp cho dự án, thành tích nổi bật) để xác định mức độ giá trị của họ đối với công ty.
  Ví dụ: Nhân viên này có dẫn dắt các dự án quan trọng hoặc đạt doanh thu vượt trội không?
- **Cơ cấu lương nội bộ**:
  Mức lương hiện tại của nhân viên so sánh với các nhân viên cùng vị trí hoặc cấp bậc trong công ty. Điều này giúp đảm bảo tính công bằng và tránh phá vỡ cấu trúc lương.
  Ví dụ: Mức lương trung bình cho vị trí tương tự tại công ty là bao nhiêu?
- **Ngân sách công ty**:
  Dữ liệu tài chính nội bộ để đánh giá khả năng chi trả cho khoản tăng lương 25 triệu VND/tháng mà không ảnh hưởng đến lợi nhuận hoặc các chi phí khác.
- **Tỷ lệ nghỉ việc**:
  Thống kê nội bộ về tỷ lệ nhân viên nghỉ việc ở vị trí tương tự để đánh giá mức độ cần thiết phải giữ chân nhân viên này.

2. Nguồn dữ liệu bên ngoài:

- **Khảo sát lương thị trường**:
  Dữ liệu từ các báo cáo lương của các công ty tuyển dụng (như VietnamWorks, Navigos, hoặc JobStreet) để so sánh mức lương hiện tại và mức lương đề xuất với thị trường lao

động Việt Nam năm 2025.
Ví dụ: Mức lương trung bình cho một kỹ sư phần mềm cấp cao tại TP.HCM là bao nhiêu?

- **Mức lương của đối thủ cạnh tranh**:
  Thông tin về chính sách lương thưởng của các công ty cùng ngành (thu thập qua mạng lưới chuyên môn hoặc báo cáo ngành) để xác định xem mức lương đề xuất có đủ sức cạnh tranh để giữ chân nhân viên.
- **Xu hướng thị trường lao động**:
  Dữ liệu về nhu cầu nhân sự trong ngành (ví dụ: ngành công nghệ, tài chính) tại Việt Nam năm 2025, bao gồm mức độ khan hiếm nhân tài ở vị trí tương tự.
  Ví dụ: Có bao nhiêu công ty đang tuyển vị trí này với mức lương cao hơn?
- **Chỉ số giá tiêu dùng (CPI)**:
  Dữ liệu từ Tổng cục Thống kê Việt Nam để đánh giá mức tăng lương phù hợp với lạm phát và chi phí sinh hoạt tại các thành phố lớn như Hà Nội hoặc TP.HCM.

Câu hỏi thảo luận:

- Làm thế nào để cân bằng giữa dữ liệu nội bộ và bên ngoài khi đưa ra quyết định?
- Nếu ngân sách công ty hạn chế, dữ liệu nào sẽ quan trọng nhất để ưu tiên?

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Bài tập: Thống kê về nguyên nhân tử vong ở Việt Nam

Trong một nghiên cứu gần đây tại Việt Nam về nguyên nhân tử vong ở nam giới từ 60 tuổi trở lên, một mẫu gồm 150 nam giới cho thấy 60 người tử vong do các bệnh liên quan đến tim mạch.

Câu hỏi:

a. Xây dựng một thống kê mô tả có thể được sử dụng để ước lượng tỷ lệ phần trăm nam giới từ 60 tuổi trở lên tử vong do các bệnh liên quan đến tim mạch.
b. Dữ liệu về nguyên nhân tử vong là định tính (categorical) hay định lượng (quantitative)?
c. Thảo luận về vai trò của suy luận thống kê trong loại nghiên cứu y học này.

Đáp án gợi ý:

a. Ước lượng tỷ lệ phần trăm:

Để ước lượng tỷ lệ phần trăm nam giới từ 60 tuổi trở lên tử vong do bệnh tim mạch, ta tính tỷ lệ phần trăm dựa trên mẫu:

- Số người tử vong do bệnh tim mạch: 60
- Tổng số người trong mẫu: 150
- Tỷ lệ phần trăm = (60 / 150) × 100 = **40%**

Vậy, khoảng **40%** nam giới từ 60 tuổi trở lên trong mẫu tử vong do các bệnh liên quan đến tim mạch. Đây là một thống kê mô tả dùng để ước lượng tỷ lệ trong tổng thể.

b. Loại dữ liệu:

Dữ liệu về nguyên nhân tử vong là **định tính (categorical)**.

- Lý do: Nguyên nhân tử vong được phân loại thành các danh mục (ví dụ: bệnh tim mạch, ung thư, tai nạn, v.v.), không phải là số đo hay giá trị số.

c. Vai trò của suy luận thống kê trong nghiên cứu y học:

Suy luận thống kê đóng vai trò quan trọng trong nghiên cứu y học, đặc biệt trong việc:

- **Ước lượng tổng thể**: Dựa trên mẫu 150 người, suy luận thống kê giúp ước lượng tỷ lệ tử vong do bệnh tim mạch trong toàn bộ dân số nam giới từ 60 tuổi trở lên ở Việt Nam (ví dụ: sử dụng khoảng tin cậy để xác định độ chính xác của tỷ lệ 40%).
- **Kiểm định giả thuyết**: Suy luận thống kê có thể được dùng để kiểm tra xem tỷ lệ tử vong do bệnh tim mạch có khác biệt đáng kể giữa các nhóm (ví dụ: nam giới ở thành thị so với nông thôn) hay không.
- **Hỗ trợ ra quyết định y tế**: Kết quả nghiên cứu giúp các cơ quan y tế Việt Nam (như Bộ Y tế) phân bổ nguồn lực, xây dựng chương trình phòng ngừa bệnh tim mạch, hoặc nâng cao nhận thức cộng đồng.
- **Đánh giá yếu tố nguy cơ**: Suy luận thống kê có thể phân tích mối liên hệ giữa bệnh tim mạch và các yếu tố như lối sống (hút thuốc, chế độ ăn), giúp định hướng chính sách y tế công cộng.

Ví dụ: Nếu nghiên cứu mở rộng, suy luận thống kê có thể sử dụng hồi quy logistic để dự đoán xác suất tử vong do bệnh tim mạch dựa trên các biến như tuổi, huyết áp, hoặc chỉ số BMI.

Câu hỏi thảo luận:

- Làm thế nào để đảm bảo mẫu 150 người đại diện cho dân số nam giới từ 60 tuổi trở lên ở Việt Nam?
- Nếu muốn mở rộng nghiên cứu để bao gồm nữ giới, suy luận thống kê sẽ được áp dụng như thế nào?

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Bài tập: Thống kê về độc giả tạp chí kinh tế tại Việt Nam

Trong một khảo sát năm 2025, 68.7% độc giả của một tạp chí kinh tế uy tín tại Việt Nam cho biết họ đã lưu trú tại khách sạn vì mục đích công việc trong 12 tháng qua, với 28.5% độc giả sử dụng hạng thương gia hoặc hạng nhất khi đi công tác.

Câu hỏi:

a. Tổng thể quan tâm trong nghiên cứu này là gì?
b. Hạng ghế trên phương tiện di chuyển (class of travel) là biến định tính (categorical) hay định lượng (quantitative)?
c. Nếu một độc giả đã lưu trú tại khách sạn vì mục đích công việc trong 12 tháng qua, đây là biến định tính hay định lượng?
d. Nghiên cứu này sử dụng dữ liệu cross-sectional hay time series?
e. Mô tả các suy luận thống kê mà tạp chí có thể thực hiện dựa trên khảo sát này.

Đáp án gợi ý:

a. Tổng thể quan tâm:

Tổng thể quan tâm là **tất cả độc giả của tạp chí kinh tế này tại Việt Nam trong năm 2025**.

b. Loại biến của hạng ghế (class of travel):

Hạng ghế trên phương tiện di chuyển là biến **định tính (categorical)**.

- Lý do: Hạng ghế được chia thành các danh mục cố định (ví dụ: phổ thông, thương gia, hạng nhất), không phải giá trị số có thể đo lường.

c. Loại biến của việc lưu trú tại khách sạn:

Việc một độc giả đã lưu trú tại khách sạn vì mục đích công việc trong 12 tháng qua là biến **định tính (categorical)**.

- Lý do: Dữ liệu này được ghi nhận dưới dạng "có" hoặc "không" (lưu trú hoặc không lưu trú), thuộc về các danh mục cố định.

d. Loại dữ liệu:

Nghiên cứu này sử dụng dữ liệu chéo(**cross-sectional)**.

- Lý do: Dữ liệu được thu thập tại một thời điểm cụ thể (năm 2025) từ một nhóm độc giả, không theo dõi sự thay đổi qua thời gian.

e. Các suy luận thống kê có thể thực hiện:

Tạp chí có thể sử dụng suy luận thống kê để:

- **Ước lượng tổng thể**: Dựa trên mẫu khảo sát, ước lượng tỷ lệ độc giả trong toàn bộ dân số độc giả tại Việt Nam đã lưu trú tại khách sạn vì công việc (khoảng 68.7%) hoặc sử dụng hạng thương gia/hạng nhất (khoảng 28.5%). Có thể sử dụng khoảng tin cậy để đánh giá độ chính xác của các tỷ lệ này.
- **So sánh nhóm**: Kiểm tra xem có sự khác biệt đáng kể về hành vi (lưu trú khách sạn hoặc chọn hạng ghế) giữa các nhóm độc giả, ví dụ: theo độ tuổi, thu nhập, hoặc khu vực sinh sống (Hà Nội, TP.HCM, Đà Nẵng).
- **Dự đoán xu hướng**: Phân tích mối liên hệ giữa việc đi công tác và lựa chọn hạng ghế để dự đoán nhu cầu dịch vụ cao cấp (như khách sạn 5 sao hoặc vé thương gia) trong ngành du lịch công tác tại Việt Nam.
- **Hỗ trợ quảng cáo**: Kết quả khảo sát có thể được dùng để thu hút các nhà quảng cáo (khách sạn, hãng hàng không) nhắm đến đối tượng độc giả có thu nhập cao, thường xuyên đi công tác.
- **Đánh giá thị trường**: Dựa trên tỷ lệ 28.5% độc giả chọn hạng thương gia/hạng nhất, tạp chí có thể suy ra nhu cầu về dịch vụ cao cấp trong ngành hàng không tại Việt Nam, từ đó cung cấp thông tin cho các đối tác kinh doanh.

Câu hỏi thảo luận:

- Làm thế nào để đảm bảo mẫu khảo sát đại diện cho toàn bộ độc giả của tạp chí tại Việt Nam?
- Nếu tạp chí muốn mở rộng khảo sát để theo dõi xu hướng qua nhiều năm (2025, 2026, 2027), loại dữ liệu nào sẽ được sử dụng?

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft