LAB-MOVIE
LABOUR MARKET OBSERVATORY IN VIETNAM UNIVERSITIES

Co-funded by the Erasmus+ Programme of the European Union

LAB-MOVIE

Quan (Khảo) sát thị trường lao động tại các Trường Đại học Việt Nam

Gói công việc 2
Chuyển giao kiến thức và phương pháp luận phân tích thị trường lao động

Kết quả 2.1
Phát triển tài liệu đào tạo

PHÂN TÍCH DỮ LIỆU

KA2 - HỢP TÁC ĐỂ ĐỔI MỚI VÀ TRAO ĐỔI CÁC THỰC HÀNH (THỰC TIỄN) TỐT
SỐ DỰ ÁN: 609653-EPP-1-2019-1-IT-EPPKA2-CBHE-JP Lab Movie

"Sự hỗ trợ của Ủy ban Châu Âu trong việc sản xuất ấn phẩm này không bao gồm sự chứng thực các nội dung, mà chỉ phản ánh quan điểm của tác giả, và Ủy ban không chịu trách nhiệm cho bất kỳ việc sử dụng nào từ thông tin có trong đó."

PHÂN TÍCH DỮ LIỆU

Những chủ đề sẽ được kiểm tra:

1. Các loại biến.
2. Mã hóa dữ liệu.
3. Ghi dữ liệu trên máy tính hỗ trợ.
4. Chỉnh sửa dữ liệu được thu thập:
   4.1. Làm sạch dữ liệu.
   4.2. Mất dữ liệu.
5. Xử lý dữ liệu:
   5.1. Bảng dữ liệu.
   5.2. Trọng số khảo sát.

1. Các loại biến

Biến là đặc điểm, thường là cơ bản, đề cập đến các đơn vị thống kê, do đó, là đặc điểm của quần thể thống kê.

Tập đáp ứng đại diện cho các kết quả có thể của một phép đo, tức là các giá trị trong đó một biến số có thể thay đổi.

Phân loại các biến là tập các giá trị mà một biến có thể nhận được (ở mức độ chi tiết tối đa).

Các biến có thể là:

- **Định lượng** (đo được bằng bằng số).
- **Định tính** (không đo được bằng bằng số).

Các biến định lượng có thể là:

- **Rời rạc** (tập hợp các giá trị mà chúng có thể giả sử là hữu hạn hoặc bằng số; tập hợp các số tự nhiên 1,2,3,4, ...).
- **Liên tục** (tập hợp các giá trị được giả định là tập các số thực hoặc một dãy số thực).

Các biến định tính có thể là:

- **Thứ tự** (tập hợp các câu trả lời có thứ tự tự nhiên, ví dụ : có thể được sắp xếp theo một thang điểm).
- **Danh định** (tập hợp các câu trả lời không có thứ tự tự nhiên).

2. Mã hóa dữ liệu

Là vấn đề về chuẩn bị dữ liệu thu thập được để sau đó được chuyển đến hệ thống máy tính hỗ trợ.

Thông qua mã hóa dữ liệu, mã số được gắn với từng giá trị phản hồi (mã hóa được thực hiện khi chuẩn bị bảng câu hỏi - xem bảng câu hỏi).

- Khi đối diện với các câu trả lời như "Tôi không biết", "không trả lời", "Tôi không nhớ", nên sử dụng một mã cụ thể áp dụng cho tất cả các câu hỏi (bạn có thể sử dụng các chữ số như 99 cho "Tôi không biết", 98 cho "không trả lời", 97 cho "Tôi không nhớ").

- Đối với các câu hỏi không áp dụng được, nên sử dụng các mã phân biệt dễ dàng như 777, 888, 999 và tương tự.

Việc sử dụng các mã này rất hữu ích vì các khoảng trống có thể gây nhầm lẫn (không rõ những gì chúng chỉ ra) và có thể đẩy các việc tính toán cho quá trình xử lý tiếp theo.

Trong trường hợp các câu hỏi mở, hoạt động như sau:

- kiểm tra câu trả lời
- quyết định cách nhóm các câu trả lời thành các danh mục đồng nhất
- liên kết mã.

Sau đó là vấn đề chuyển đổi các câu hỏi mở thành câu hỏi đóng.

Một ví dụ về mã hóa là sự hoán vị thông tin liên quan đến lĩnh vực hoạt động kinh tế của các doanh nghiệp, được người trả lời mô tả dưới dạng thông tục, thành các mã tương ứng theo phân loại ATECO.

3. Ghi dữ liệu về máy tính hỗ trợ

Là vấn đề chuyển dữ liệu mã hóa sang máy tính hỗ trợ phù hợp cho việc phân tích thống kê tiếp theo.

Mục đích chính là xây dựng ma trận dữ liệu

- chứa nhiều hàng (bản ghi) các đơn vị phân tích (ví dụ: các công ty được phỏng vấn),
- chứa nhiều cột (trường) như các biến được xem xét (là các câu hỏi được báo cáo trong bảng câu hỏi).

Một số kỹ thuật khảo sát cho thấy trước việc đồng thời nhập dữ liệu vào máy tính hỗ trợ (CATI, CAPI, CAWI).

4. Chỉnh sửa dữ liệu được thu thập

Các tình huống cần kiểm tra:

- **Giá trị ngoài miền** : các giá trị của biến không thuộc về tập hợp được xác định trước (giá trị ngoài miền làm phát sinh lỗi).
- **Giá trị bất thường** : một đơn vị là bất thường (ngoại lệ) khi nó có các đặc điểm khác biệt đáng kể so với hầu hết các đơn vị khác (các giá trị bất thường làm phát sinh các lỗi).
- **Không tương thích giữa các phản hồi** : các giá trị của một hoặc nhiều biến trái ngược với các quy tắc logic được xác định trước và/hoặc các mối quan hệ toán học (không tương thích dẫn đến các tình huống lỗi, nhưng thường không biết đến biến nào).

Loại kiểm tra:

- **Kiểm tra tính hợp lệ hoặc phạm vi** : kiểm tra xem các giá trị được giả định bởi một biến đã cho có nằm trong phạm vi định nghĩa của biến không.
- **Kiểm tra thống kê** : được sử dụng để cô lập các đơn vị có giá trị cho một số biến có trong chúng làm sai lệch đáng kể so với các giá trị được giả định bởi cùng một biến trong các đơn vị còn lại. Các giá trị này có xác suất lỗi cao, nhưng cần kiểm tra thêm.
- **Kiểm tra tính nhất quán** : kiểm tra xem các điều kiện xác định trước của các giá trị được giả định bởi các biến được lấy trong cùng một đơn vị có đáp ứng các yêu cầu nhất định hay không (quy tắc không tương thích).

Quy hoạch không tương thích

Kiểm tra tính nhất quán được sử dụng để xây dựng quy hoạch không tương thích.

Một quy hoạch không tương thích là một tập hợp các ràng buộc không dư thừa và không mâu thuẫn phải được đáp ứng đồng thời bởi mỗi đơn vị thống kê để thông tin tương ứng được coi là chính xác.

Các quy tắc tạo nên một quy hoạch không tương thích có thể được phân biệt thành:

- **Quy tắc chính thức** , bắt nguồn từ cấu trúc của bảng câu hỏi, ví dụ: trực tiếp từ các quy tắc biên dịch và các đường dẫn nội bộ của bảng câu hỏi.
- **Các quy tắc về bản chất** xuất phát từ các xem xét thống kê/toán học, hoặc từ một kiến thức cụ thể tiên nghiệm về hiện tượng đang nghiên cứu.

Khi các bản ghi có giá trị vi phạm một hoặc nhiều ràng buộc của quy hoạch không tương thích đã được xác định, bài toán sẽ trở thành xác định các biến chịu trách nhiệm cho vi phạm này, thực tế là các biến mà giá trị của chúng chỉ được coi là không đúng (mất mát) và đúng.

4.1. Làm sạch dữ liệu

Nếu tìm thấy các bản ghi chứa giá trị không chính xác và biết biến nào chịu trách nhiệm cho sự không chính xác này, có thể thực hiện các thay đổi sau:

- Trở lại nguồn.
- Loại bỏ xác định hoặc xác suất.
- Xử lý dữ liệu như một đối tượng không phản hồi.

Trở lại nguồn

Nếu có thể, hãy yêu cầu dữ liệu trực tiếp từ người trả lời hoặc người đã hoàn thành bảng câu hỏi (rất tốn thời gian và công sức, nhưng đảm bảo rằng dữ liệu chính xác nhất)

Loại bỏ xác định

Là gán giá trị cho biến thay vì giá trị sai, dựa trên thông tin khác (ví dụ: nếu tuổi <14 thì số trẻ em = 0).

Thông thường, thông tin phải có sẵn trong tập dữ liệu và bản thân đơn vị.

Nói chung, để tiến hành loại bỏ một cách chắc chắn, thông tin cần đến từ nhiều hơn một biến, nếu không sẽ không thể quyết định (giữa hai biến có giá trị không tương thích) biến nào là đúng.

Loại bỏ xác suất

Điều đó có nghĩa là gán giá trị cho biến thay cho giá trị sai, dựa trên những những quan sát trong các đơn vị có cùng đặc điểm.

Xử lý dữ liệu như một đối tượng không phản hồi

Nếu không có công cụ để khôi phục hoặc gán thuộc tính dữ liệu đúng, tất cả những gì còn lại là xóa dữ liệu sai và coi trường đó thiếu dữ liệu.

4.2. Mất dữ liệu

Một vấn đề rất thường xuyên trong các cuộc khảo sát mẫu là không có phản hồi.

Không phản hồi có thể là:

- **Hoàn toàn** (một hoặc nhiều đơn vị không trả lời toàn bộ bảng câu hỏi).
- **Một phần** (một hoặc nhiều đơn vị không trả lời một hoặc nhiều câu hỏi).

Hoàn toàn không phản hồi

Vì lý do tự nguyện ("người trả lời" không muốn trả lời) hoặc vô tình (vắng mặt vì công việc, vì cuộc khảo sát được thực hiện ở trường, v.v.), luôn có một hạn mức, lớn hơn hoặc ít hơn của những người nhưng không phản hồi, mặc dù là một phần của mẫu được xác định.

Nói chung, trong những trường hợp này, việc thay thế được thực hiện bằng cách chọn những người mới có cùng đặc điểm với những người đã từ chối (tổng tỷ lệ không phản hồi lên đến 20% có thể được coi là chấp nhận được).

Tuy nhiên thích hợp với việc

- thực hiện kiểm soát chặt chẽ mẫu thu được (để đảm bảo rằng mẫu thực sự "tương tự" với quần thể).
- kiểm tra tính đồng nhất với quần thể đối với một số biến liên quan không có trong kế hoạch lấy mẫu

Không phản hồi một phần

Cũng luôn có một hạn mức của những người vì nhiều lý do khác nhau không trả lời tất cả các câu hỏi (họ không biết hoặc không muốn trả lời).

Trong những trường hợp này, với sự quan tâm đúng mức và chỉ trong những trường hợp rõ ràng/rõ ràng nhất, có thể gán giá trị cho biến, thay vì giá trị bị thiếu, thông qua loại bỏ xác định. Không phản hồi một phần có thể được bao gồm trong kiểm tra tính nhất quán (quy hoạch không tương thích).

5. Xử lý dữ liệu

Xử lý dữ liệu cho phép cung cấp thông tin thống kê.
Cung cấp thông tin thống kê là:

- Thiết lập hệ thống trình bày và tham vấn dữ liệu phù hợp với các bên liên quan và hoàn hảo về mặt phương pháp luận (dù trên Internet hay trên giấy).
- Để xác định và cấu trúc thông tin thống kê bằng hệ thống tiêu chí phù hợp.
- Chuẩn bị dữ liệu thống kê (bảng và biểu đồ) một cách chính xác và dễ đọc.

Trong số các hình thức khác nhau, thể hiện bằng hình ảnh và bảng là đặc biệt quan trọng (là những thể hiện thực sự được quan tâm, phân tích mô tả).

5.1. Bảng dữ liệu

Bảng dữ liệu mô tả một biến hoặc mối quan hệ giữa hai hoặc nhiều biến.

Bảng phải chứa tất cả thông tin cần thiết để hiểu dữ liệu, bất kể dạng văn bản được chèn vào (bảng hỗ trợ văn bản, phải rõ ràng ngay cả khi không có bảng). Nó bao gồm:

- Tiêu đề.
- Thân.

Dữ liệu trong bảng có thể được thể hiện nhiều hơn bằng biểu đồ. Biểu đồ chỉ có thể thay thế bảng nếu nó chứa cùng thông tin.

5.1.1. Tiêu đề bảng

Tiêu đề của bảng phải được giải thích một cách hoàn hảo:

- Những gì chúng ta đang nói đến (công ty, dân số, người dùng, người dân, con người, ...), do đó, xác định rõ đơn vị được trình bày trong bảng.
- Chỉ định loại dữ liệu nào hiện có trong bảng (giá trị tuyệt đối, tỷ lệ phần trăm, tỷ lệ, ...).
- Chỉ định biến phụ và chính.
- Chỉ định vị trí, khoảng thời gian tham chiếu (nếu bảng được xây dựng từ dữ liệu hành chính hoặc các nguồn chính thức, thì phải trích dẫn nguồn).

Bảng có thể hiển thị số (giá trị tuyệt đối hoặc tỷ lệ) hoặc tỷ lệ phần trăm.

5.1.2. Bảng số

Tiêu đề: cái gì, mô tả với loại dữ liệu nào, theo các biến nào. Địa điểm, năm.

	Header variable	
Side variable	Table body (data)	Row totals
	Column totals	Total

Bảng có giá trị tuyệt đối luôn phải có tổng số hàng và cột.

5.1.3. Bảng tỷ lệ phần trăm

Bảng có thể chứa:

- Phần trăm cột (là những phần được quan tâm).
- Phần trăm hàng.
- Phần trăm ô.

Các bảng có phân bố tỷ lệ phần trăm phải được đánh dấu 100. Nếu phân bố là trên mỗi hàng hoặc cột, thì cũng phải có phân bố tỷ lệ phần trăm cận biên, không phải là tổng của tỷ lệ phần trăm.

Ví dụ về bảng có phần trăm cột.

Tiêu đề: cái gì, được mô tả theo tỷ lệ phần trăm, theo các biến nào. Địa điểm, năm.

	Header variable (independent)				
Side variable (dependent)	x	x	x	x	x
	x	x	x	x	x
	x	x	x	x	x
	x	x	x	x	x
	100	100	100	100	100

Trường hợp đặc biệt

Trong trường hợp các câu hỏi có nhiều lựa chọn, phải chú ý đến việc trình bày tỷ lệ phần trăm: tổng của chúng vượt quá 100, do đó phải ghi chú giải thích và không được thêm tổng số phần trăm.

5.1.4. Thiếu dữ liệu trong bảng

Trong bảng số tuyệt đối: dữ liệu bị thiếu được coi là một giá trị và phải được nhập vào hàng cuối cùng và cột cuối cùng, trước tổng số. Loại trừ dữ liệu bị thiếu khỏi bảng số tuyệt đối có nghĩa là trình bày các bảng với các tổng số khác nhau.

Trong các bảng có tỷ lệ phần trăm:

1. Chèn phần trăm dữ liệu bị thiếu vào bên trong bảng, coi chúng như một giá trị (có thể không có ý nghĩa gì nếu chèn phần trăm dữ liệu bị thiếu trong bảng, bởi vì theo cách này, chúng ta không biết phân phối thực của hiện tượng).
2. Tính toán phần trăm phân phối sau khi loại trừ dữ liệu bị thiếu, sẽ được chỉ ra ở cuối bảng.
3. Phân bố lại dữ liệu bị thiếu trong bảng, tôn trọng trường hợp phân bố trước đó (chấp nhận được nếu dữ liệu bị thiếu được phân bổ).

5.2. Trọng số khảo sát

- Mỗi đơn vị trong mẫu đại diện cho một số đơn vị nhất định trong quần thể, sao cho toàn bộ mẫu đại diện cho toàn bộ quần thể.
- Khi chúng tôi cung cấp kết quả khảo sát, mỗi đơn vị phải được liên kết với trọng số của nó, sao cho tổng số đơn vị tương ứng với N.
- Trọng số liên quan đến mỗi đơn vị được gọi là trọng số khảo sát và bằng với tỷ lệ nghịch của xác suất lựa chọn.

Sử dụng trọng số khảo sát

- Trong **các mẫu trọng số tự có**, trọng số không đổi cho tất cả các đơn vị mẫu. Các ước tính vẫn giữ nguyên có hoặc không có trọng số (nhưng phương sai ước tính hoặc thay đổi lỗi mẫu).
- Trong **mẫu trọng số không tự có**, việc sử dụng trọng số là cần thiết để có được ước tính đúng.