

DH2.21

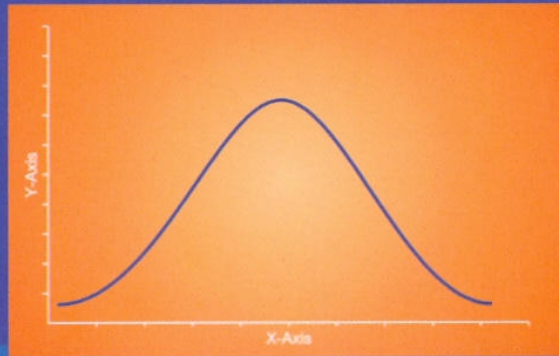
BỘ Y TẾ

THỐNG KÊ Y TẾ CÔNG CỘNG

Phân Phân Tích Số Liệu

SÁCH DÙNG ĐÀO TẠO CỬ NHÂN Y TẾ CÔNG CỘNG

Chủ biên: TS. Phạm Việt Cường



THƯ VIỆN
HUBT

NHÀ XUẤT BẢN Y HỌC
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

BỘ Y TẾ

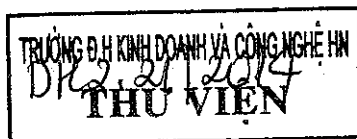
THỐNG KÊ Y TẾ CÔNG CỘNG

PHẦN PHÂN TÍCH SỐ LIỆU

(SÁCH ĐÀO TẠO CỬ NHÂN Y TẾ CÔNG CỘNG)

Mã số: Đ.14.Y.12

Chủ biên: TS. Phạm Việt Cường



CHỈ ĐẠO BIÊN SOẠN:

Vụ Khoa học & Đào tạo, Bộ Y tế

CHỦ BIÊN:

TS. Phạm Việt Cường

THAM GIA BIÊN SOẠN:

TS. Phạm Việt Cường

TS. Lê Cự Linh

Ths. Lê Thanh Hà

Ths. Bùi Tú Quyên

TS. Diana Battistutta

Ths. Kate Halton

THƯ KÝ BIÊN SOẠN:

Th.S. Nguyễn Trang Nhung

Th.S. Nguyễn Thị Hồng

THAM GIA TỔ CHỨC BẢN THẢO:

ThS. Phí Văn Thâm

ThS. Phí Nguyệt Thanh

© Bản quyền thuộc Bộ Y tế (Vụ Khoa học và Đào tạo)

LỜI GIỚI THIỆU

Thực hiện một số điều của Luật Giáo dục, Bộ Giáo dục & Đào tạo và Bộ Y tế đã ban hành chương trình khung đào tạo Cử nhân Y tế Công cộng. Bộ Y tế tổ chức biên soạn tài liệu dạy – học các môn học chuyên môn, cơ bản chuyên ngành theo chương trình trên nhằm từng bước xây dựng bộ tài liệu dạy – học chuẩn về chuyên môn để đảm bảo chất lượng đào tạo nhân lực y tế.

Sách “*Thống kê y tế công cộng*” được biên soạn dựa trên chương trình giáo dục đại học của Trường Đại học Y tế Công cộng trên cơ sở chương trình khung đã được phê duyệt. Sách được biên soạn với phương châm: cung cấp kiến thức cơ bản, có hệ thống; nội dung chính xác, khoa học, cập nhật các tiến bộ khoa học, kỹ thuật hiện đại và thực tiễn của Việt Nam.

Sách “*Thống kê y tế công cộng*” đã được biên soạn bởi các nhà giáo giàu kinh nghiệm và tâm huyết của bộ môn Thống kê Trường Đại học Y tế Công cộng. Sách “*Thống kê y tế công cộng*” đã được hội đồng chuyên môn thẩm định sách và tài liệu dạy – học chuyên ngành Cử nhân Y tế Công cộng của Bộ Y tế thẩm định vào năm 2009. Bộ Y tế ban hành là tài liệu dạy – học đạt chuẩn chuyên môn của ngành y tế trong giai đoạn hiện nay. Trong quá trình sử dụng sách phải được chỉnh lý, bổ sung và cập nhật.

Bộ Y tế xin chân thành cảm ơn các nhà giáo, các chuyên gia của Trường Đại học Y tế Công cộng đã dành nhiều công sức hoàn thành cuốn sách, TS. Đỗ Văn Dũng và ThS. Lưu Ngọc Hoạch đã đọc và phản biện để cuốn sách sớm hoàn thành kịp thời phục vụ cho công tác đào tạo nhân lực y tế. Vì lần đầu xuất bản, chúng tôi mong nhận được ý kiến đóng góp của đồng nghiệp, các bạn sinh viên và các độc giả để lần xuất bản sau được hoàn thiện hơn.

**VỤ KHOA HỌC VÀ ĐÀO TẠO
BỘ Y TẾ**





**THƯ VIỆN
HUBT**

TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

LỜI NÓI ĐẦU

Cuốn giáo trình này nằm trong bộ giáo trình Thống kê Y tế Công cộng và được biên soạn với mục tiêu là phát triển kỹ năng thực hành và áp dụng kiến thức thống kê trong việc xử lý và phân tích số liệu. Cuốn giáo trình này sẽ giúp người học ôn tập và thực hành các phương pháp thống kê đã được học trong chương trình Thống kê Y tế Công cộng phần Thống kê cơ bản, đồng thời cũng cung cấp thêm các phương pháp phân tích thống kê nâng cao hơn.

Giáo trình bao gồm 6 bài học được liên kết rất chặt chẽ, các khái niệm của bài trước sẽ cần thiết để hiểu các khái niệm của các phần sau. Tất cả các phương pháp thống kê và các kỹ năng phân tích trong giáo trình đều được trình bày và thực hiện phân tích bằng phần mềm SPSS và dựa trên một bộ số liệu từ một cuộc điều tra tại Việt Nam. Các nội dung trong giáo trình được trình bày theo trình tự từ đơn giản đến phức tạp, do đó giáo trình này có thể sử dụng để giảng dạy cho nhiều đối tượng học viên khác nhau (cả trình độ đại học và sau đại học), những nhà nghiên cứu, cán bộ làm việc trong các lĩnh vực khoa học sức khoẻ nói chung. Việc thiết kế theo trình tự và hướng dẫn cụ thể từng bước còn cho phép người đọc tự học tập và tra cứu khi sử dụng thống kê và phần mềm máy tính để phân tích một bộ số liệu.

Đây là một giáo trình được xây dựng theo một hình thức mới, lồng ghép việc trình bày lý thuyết và hướng dẫn kỹ năng thực hành dựa trên ví dụ thực tiễn ở Việt Nam. Nhóm tác giả của giáo trình này là những giảng viên có chuyên ngành thống kê, dân số và dịch tễ học do đó nội dung của giáo trình được thể hiện lồng ghép chứ không chỉ là một giáo trình toán thống kê. Nhóm tác giả cũng gửi lời cảm ơn tới Tiến sỹ Diana Battistutta và Thạc sỹ Kate Halton, Trường Y tế Công cộng, Đại học Công nghệ Queensland, Australia đã có những đóng góp rất lớn trong việc xây dựng và hoàn thiện cuốn sách này.

Mặc dù là tài liệu này đã được chỉnh sửa và cập nhật nhưng cuốn sách này cũng không tránh khỏi những khiếm khuyết. Chúng tôi mong rằng học viên, giảng viên cũng như bạn đọc trong và ngoài ngành sẽ đưa ra nhiều ý kiến đóng góp quý báu để giáo trình ngày càng hoàn thiện và phục vụ người học tốt hơn.

Hà Nội, tháng 6 năm 2009

Các tác giả





MỤC LỤC

Giới thiệu các phương pháp nghiên cứu định lượng	9
1.1. Các bước tiến hành nghiên cứu	10
1.2. Câu hỏi và giả thuyết nghiên cứu	12
1.3. Các thiết kế nghiên cứu định lượng cơ bản.....	14
1.4. Các thành phần của thiết kế có ảnh hưởng tới việc phân tích.....	17
Quản lý số liệu	21
2.1. Bộ số liệu mẫu.....	21
2.2. Xử lý thông tin nghiên cứu cho phân tích định lượng.....	22
2.3. Các ví dụ về làm sạch số liệu.....	32
Phân tích thống kê mô tả	59
3.1. Giới thiệu	59
3.2. Tiến trình của kế hoạch phân tích	60
3.3. Các câu hỏi nghiên cứu từ bộ số liệu mẫu.....	61
3.4. Kế hoạch phân tích của bộ số liệu mẫu – thống kê mô tả.....	63
3.5. Phân tích mô tả cho một biến.....	66
3.6. Mô tả các mối liên quan.....	75
3.7. Viết kết quả của phân tích mô tả	84
Phân tích thống kê suy luận	88
4.1. Giới thiệu	88
4.2. Quá trình lập kế hoạch phân tích số liệu	89
4.3. Giả thuyết thống kê.....	90
4.4. Sử dụng kiểm định nào?	90
4.5. Sử dụng SPSS để kiểm định giả thuyết.....	91
4.6. Trình bày kết quả của các phân tích suy luận	145
4.7. Giả định	147
Nhiều và không chế nhiều	161
5.1. Giới thiệu	161
5.2. Nhiều	161
Tính cỡ mẫu	171
6.1. Các yếu tố ảnh hưởng đến tính tin cậy trong các kết quả	171
6.2. Những điều kiện cần thiết để tính cỡ mẫu.....	175
6.3. Tính cỡ mẫu	176



MỤC TIÊU KHOÁ HỌC

Sau khi hoàn thành khoá học này, học viên có thể:

1. Chọn được loại kiểm định thống kê phù hợp cho các loại câu hỏi nghiên cứu khác nhau về sức khoẻ.
2. Phiên giải được kết quả phân tích từ phần mềm SPSS.
3. Viết được các báo cáo cho các kết quả phân tích thống kê.
4. Sử dụng được phần mềm SPSS để quản lý số liệu.
5. Sử dụng phần mềm SSize để tính cỡ mẫu cho các loại câu hỏi nghiên cứu khác nhau.

Bài số

1

GIỚI THIỆU CÁC PHƯƠNG PHÁP NGHIÊN CỨU ĐỊNH LƯỢNG

**“Để phân tích một bộ số liệu
bạn phải hiểu được thiết kế nghiên cứu đó”**

Thông điệp trên sẽ xuyên suốt toàn bộ môn học này. Bạn sẽ không thể có được các kết quả phân tích số liệu đúng nếu như bạn không hiểu rõ về câu hỏi nghiên cứu, không nắm được các số liệu đã được thu thập như thế nào. Những chiến lược được dùng để có được những số liệu cần thiết được gọi là thiết kế nghiên cứu. Bạn sẽ không thể thực hiện được các phép phân tích thống kê chính xác nếu bạn không có kiến thức và hiểu biết về thiết kế nghiên cứu đã được tiến hành.

Mục tiêu

Bài đầu tiên của khoá học này sẽ cung cấp cho bạn một cách tóm tắt về tổng quan thiết kế nghiên cứu trong Y tế Công cộng, tuy nhiên chúng tôi sẽ không đi sâu vào các chi tiết thiết kế, bạn đọc có thể tìm đọc những nội dung này trong các sách về nghiên cứu hệ thống y tế hoặc Dịch tễ học. Bài này sẽ tập trung vào các khái niệm rất quan trọng mà bạn cần phải lưu tâm khi tiến hành phân tích một bộ số liệu bất kỳ, các khái niệm đó bao gồm:

1. Các giả thuyết nghiên cứu và các giả thuyết thống kê.
2. Loại thiết kế (nghiên cứu thực nghiệm hay nghiên cứu quan sát).
3. Định nghĩa đơn vị quan sát và các quan sát (vd: sự phụ thuộc trong dữ liệu).
4. Xác định các nguồn sai số tiềm tàng (nhiều, và các sai số do chọn mẫu).
5. Ai sẽ là người sử dụng các báo cáo của bạn (báo cáo khoa học hay báo cáo dành cho những đối tượng người đọc khác).

1.1. Các bước tiến hành nghiên cứu

Nghiên cứu thường được tiến hành do những nhận thức hiện có của chúng ta (hoặc mức độ hiểu biết của chúng ta) về một vấn đề nào đó (đôi khi được gọi là “sự thực”) được cho là không đúng hoặc chưa đầy đủ. Một nhà nghiên cứu thường đưa ra một giả thuyết rằng có một quan điểm khác dường như có thể được coi là đúng đắn hơn và mục đích cho việc thu thập số liệu để chứng minh giả thuyết đó. Nếu những số liệu thu thập được phù hợp với giả thuyết của nhà nghiên cứu mới đưa ra thì có nghĩa là nhà nghiên cứu đã đúng khi nghi ngờ “sự thực” trước kia. Vậy nghiên cứu là một quá trình thu thập các bằng chứng để ủng hộ hoặc bác bỏ một quan điểm nào đó. Quan điểm của nhà nghiên cứu chính là *đổi thuyết* (alternative hypothesis) và “sự thực” đã biết chính là *giả thuyết không* (thường được gọi tắt là *giả thuyết* – null hypothesis). Bằng chứng chính là các dữ liệu, và việc khẳng định hay bác bỏ “sự thực” hiện nay hay chấp nhận “sự thực” mới do nhà nghiên cứu đưa ra (còn gọi là *đổi thuyết*) chính là các kiểm định thống kê.

Mục đích của nghiên cứu là thu thập các thông tin chính xác nhất có thể với nguồn lực hiện có, với mục tiêu cung cấp các bằng chứng chính xác để trả lời câu hỏi của nhà nghiên cứu. **Thiết kế nghiên cứu** giống như là một bài tập quản lý, nó bao gồm việc lập kế hoạch cho quá trình thu thập thông tin sao cho tiết kiệm nguồn lực (thời gian, tài chính và nhân lực). Quá trình nghiên cứu bao gồm ba bước chính:

1. Thiết kế nghiên cứu.
2. Thu thập số liệu.
3. Phân tích và phiên giải số liệu.

Giáo trình này tập trung vào bước thứ 3, tuy nhiên việc phân tích thống kê sẽ không thể tiến hành được nếu thiếu các kiến thức về hai bước đầu.

Phần lớn các câu hỏi nghiên cứu sẽ tập trung vào đánh giá sự khác biệt giữa các nhóm hoặc sự khác nhau trong một nhóm qua thời gian. Chúng ta sẽ quan tâm đến sự *biến thiên* giữa các nhóm hoặc qua các giai đoạn thời gian trong một nhóm. Càng nhiều nguồn biến thiên khác nhau của các thông tin thu thập được thì càng có nhiều cách giải thích kết quả nghiên cứu của chúng ta. Một phương pháp nghiên cứu tốt liên quan đến việc kiểm soát được các nguồn biến thiên có thể có. Hai nguồn biến thiên chính của số liệu là sự biến thiên do sự khác nhau giữa các cá thể và sự biến thiên do việc đo lường. Chúng ta không thể kiểm soát hết các biến thiên này khi chúng ta tiến hành nghiên cứu trên thực tế, do vậy chúng ta có những sai sót (trong việc chọn sai đối tượng nghiên cứu, trong việc đo lường các biến, v.v.). Bất kỳ một lỗi nào chúng ta mắc phải đều ảnh hưởng đến những mức độ sai lệch kết quả nghiên cứu của chúng ta. Ngoài ra, những

Bài số

1

GIỚI THIỆU CÁC PHƯƠNG PHÁP NGHIÊN CỨU ĐỊNH LƯỢNG

**“Để phân tích một bộ số liệu
bạn phải hiểu được thiết kế nghiên cứu đó”**

Thông điệp trên sẽ xuyên suốt toàn bộ môn học này. Bạn sẽ không thể có được các kết quả phân tích số liệu đúng nếu như bạn không hiểu rõ về câu hỏi nghiên cứu, không nắm được các số liệu đã được thu thập như thế nào. Những chiến lược được dùng để có được những số liệu cần thiết được gọi là thiết kế nghiên cứu. Bạn sẽ không thể thực hiện được các phép phân tích thống kê chính xác nếu bạn không có kiến thức và hiểu biết về thiết kế nghiên cứu đã được tiến hành.

Mục tiêu

Bài đầu tiên của khoá học này sẽ cung cấp cho bạn một cách tóm tắt về tổng quan thiết kế nghiên cứu trong Y tế Công cộng, tuy nhiên chúng tôi sẽ không đi sâu vào các chi tiết thiết kế, bạn đọc có thể tìm đọc những nội dung này trong các sách về nghiên cứu hệ thống y tế hoặc Dịch tễ học. Bài này sẽ tập trung vào các khái niệm rất quan trọng mà bạn cần phải lưu tâm khi tiến hành phân tích một bộ số liệu bất kỳ, các khái niệm đó bao gồm:

1. Các giả thuyết nghiên cứu và các giả thuyết thống kê.
2. Loại thiết kế (nghiên cứu thực nghiệm hay nghiên cứu quan sát).
3. Định nghĩa đơn vị quan sát và các quan sát (vd: sự phụ thuộc trong dữ liệu).
4. Xác định các nguồn sai số tiềm tàng (nhiều, và các sai số do chọn mẫu).
5. Ai sẽ là người sử dụng các báo cáo của bạn (báo cáo khoa học hay báo cáo dành cho những đối tượng người đọc khác).

1.1. Các bước tiến hành nghiên cứu

Nghiên cứu thường được tiến hành do những nhận thức hiện có của chúng ta (hoặc mức độ hiểu biết của chúng ta) về một vấn đề nào đó (đôi khi được gọi là “sự thực”) được cho là không đúng hoặc chưa đầy đủ. Một nhà nghiên cứu thường đưa ra một giả thuyết rằng có một quan điểm khác dường như có thể được coi là đúng đắn hơn và mục đích cho việc thu thập số liệu để chứng minh giả thuyết đó. Nếu những số liệu thu thập được phù hợp với giả thuyết của nhà nghiên cứu mới đưa ra thì có nghĩa là nhà nghiên cứu đã đúng khi nghi ngờ “sự thực” trước kia. Vậy nghiên cứu là một quá trình thu thập các bằng chứng để ủng hộ hoặc bác bỏ một quan điểm nào đó. Quan điểm của nhà nghiên cứu chính là *đối thuyết* (alternative hypothesis) và “sự thực” đã biết chính là *giả thuyết không* (thường được gọi tắt là *giả thuyết* – null hypothesis). Bằng chứng chính là các dữ liệu, và việc khẳng định hay bác bỏ “sự thực” hiện nay hay chấp nhận “sự thực” mới do nhà nghiên cứu đưa ra (còn gọi là đối thuyết) chính là các kiểm định thống kê.

Mục đích của nghiên cứu là thu thập các thông tin chính xác nhất có thể với nguồn lực hiện có, với mục tiêu cung cấp các bằng chứng chính xác để trả lời câu hỏi của nhà nghiên cứu. **Thiết kế nghiên cứu** giống như là một bài tập quản lý, nó bao gồm việc lập kế hoạch cho quá trình thu thập thông tin sao cho tiết kiệm nguồn lực (thời gian, tài chính và nhân lực). Quá trình nghiên cứu bao gồm ba bước chính:

1. Thiết kế nghiên cứu.
2. Thu thập số liệu.
3. Phân tích và phiên giải số liệu.

Giáo trình này tập trung vào bước thứ 3, tuy nhiên việc phân tích thống kê sẽ không thể tiến hành được nếu thiếu các kiến thức về hai bước đầu.

Phần lớn các câu hỏi nghiên cứu sẽ tập trung vào đánh giá sự khác biệt giữa các nhóm hoặc sự khác nhau trong một nhóm qua thời gian. Chúng ta sẽ quan tâm đến sự *biến thiên* giữa các nhóm hoặc qua các giai đoạn thời gian trong một nhóm. Càng nhiều nguồn biến thiên khác nhau của các thông tin thu thập được thì càng có nhiều cách giải thích kết quả nghiên cứu của chúng ta. Một phương pháp nghiên cứu tốt liên quan đến việc kiểm soát được các nguồn biến thiên có thể có. Hai nguồn biến thiên chính của số liệu là sự biến thiên do sự khác nhau giữa các cá thể và sự biến thiên do việc đo lường. Chúng ta không thể kiểm soát hết các biến thiên này khi chúng ta tiến hành nghiên cứu trên thực tế, do vậy chúng ta có những sai sót (trong việc chọn sai đối tượng nghiên cứu, trong việc đo lường các biến, v.v.). Bất kỳ một lỗi nào chúng ta mắc phải đều ảnh hưởng đến những mức độ sai lệch kết quả nghiên cứu của chúng ta. Ngoài ra, những

kết quả nghiên cứu chúng ta có được trong ngày hôm nay có thể khác với các kết quả nghiên cứu của những ngày khác do việc một đối tượng nghiên cứu rất có thể sẽ đưa ra những câu trả lời khác nhau ở những thời điểm khác nhau với cùng một câu hỏi. Mục đích của một nghiên cứu tốt là cố gắng giảm tối đa các nguồn có thể gây sai số. Biện pháp chính là kiểm soát nhiều nguồn sai số nhất có thể được (ngoài những biến thiên của các cá thể mà chúng ta khó có thể kiểm soát được).

Có hai loại sai số: *sai số ngẫu nhiên (random error)* và *sai số hệ thống (systematic error, hay bias)*. Sai số ngẫu nhiên có thể được định nghĩa là một thành phần không thể dự đoán được. Sai số hệ thống là sai số do đo lường dẫn tới các kết quả nghiên cứu có sự sai lệch một cách có hệ thống. Thông thường, một tính chất hoặc đặc tính có thể được đo lường mà không gặp sai số hệ thống một cách tổng thể nhưng lại có những sai số ngẫu nhiên khác nhau trong các nhóm nhỏ, hoặc trên một số đối tượng nghiên cứu.

Sai số đo lường một cách hệ thống được coi là nghiêm trọng hơn là các sai số ngẫu nhiên. Sai số ngẫu nhiên dẫn tới sự thiếu chính xác, có nghĩa thông thường là làm biến mất hoặc mờ đi sự khác biệt giữa các nhóm. Trái lại, sự nghiêm trọng của sai số hệ thống là có thể làm cho sự khác biệt giữa các nhóm bị lệch lạc và các kết luận nghiên cứu có thể bị phiến giải hoàn toàn sai lệch. Do vậy, kết quả sẽ là không chính xác (inaccurate) và không có giá trị (invalid).

Sau đây là một vài nguồn sai số hệ thống của nghiên cứu, đặc biệt là một số nguồn biến thiên quan trọng (sai số tiềm tàng) có thể ảnh hưởng đến nghiên cứu sức khỏe là:

1. *Sai số lựa chọn (selection bias)*: sai số này dẫn đến việc các nhóm được chọn lựa không đại diện được cho nhóm người mà chúng ta nghiên cứu. Điều này sẽ làm lệch lạc sự phiến giải kết quả của chúng ta (tính khái quát hoá – generalisability).
2. *Nhiều (confounding)*: sai số này xuất hiện khi so sánh các nhóm với các đặc tính khác nhau. Một biến nhiễu điển hình thường được nhắc đến là *tuổi*. Nghiên cứu thực nghiệm thường phân các đối tượng một cách ngẫu nhiên vào trong các nhóm khác nhau, cho nên tránh được nhiễu (vì các đặc tính sẽ tương đồng trong tất cả các nhóm).
3. *Sai số thông tin (information bias)*: Khi chúng ta dùng các độ đo khác nhau ở các nhóm so sánh. Ví dụ: điều tra viên có thể hỏi các câu hỏi một cách kỹ lưỡng có chủ định đối với những người bị nhiễm HIV hơn là với những người không bị nhiễm HIV khi họ biết về tình trạng nhiễm HIV của các đối tượng.



Có rất nhiều nguồn sai số khác nhau trong nghiên cứu, trên đây là những loại sai số chính cần để ý khi khi phân giải các kết quả nghiên cứu của bạn. Là người phân tích số liệu, công việc của bạn là xác định và nếu có thể thì chỉ ra độ lớn của các nguồn sai số càng nhiều càng tốt trong phạm vi số liệu cho phép.

1.2. Câu hỏi và giả thuyết nghiên cứu

Các câu hỏi nghiên cứu thông thường được phát biểu một cách rất khái quát, và khó có thể tiến hành phân tích thống kê được cho đến khi được phân tách thành các giả thuyết khoa học có thể kiểm định được. Chủ đề nghiên cứu có thể rất rộng, chẳng hạn “sức khỏe của công nhân nhà máy sản xuất gạch men”, từ đó hàng trăm câu hỏi nghiên cứu có thể được đặt ra (ví dụ “có sự khác nhau về sức khỏe đường hô hấp của nhóm thợ làm ở mỏ đá và nhóm thợ làm ở bộ phận lò hay không?”). Đây là một câu hỏi nghiên cứu chi tiết, chỉ xem xét một vấn đề cụ thể là tình trạng hệ hô hấp của công nhân nhà máy gạch – vậy còn vấn đề sức khỏe tâm thần (sang chấn tinh thần, trầm cảm) hay các vấn đề khác như: tim mạch...? Việc chọn câu hỏi nghiên cứu trong khuôn khổ một vấn đề lớn hoàn toàn phụ thuộc vào sự ưu tiên của nhà nghiên cứu.

Định nghĩa của câu hỏi nghiên cứu cần được tiến hành cẩn thận trước khi thiết kế nghiên cứu xác định cụ thể. Câu hỏi nghiên cứu cần được chuyển sang dạng các *giả thuyết khoa học*. Nó bao gồm việc xác định các biến độc lập và biến phụ thuộc sẽ được đo lường như thế nào? Và làm thế nào để phân giải mối quan hệ của chúng.

Trong ví dụ ở trên, biến phụ thuộc là tình trạng hệ hô hấp. Như vậy vẫn còn quá rộng, liệu có thể là *dung tích thở ra gắng sức* đo bằng lít (FEV) hay là tiền sử của bệnh viêm phế quản (có mắc hay không mắc), hay là một vấn đề gì khác. Như vậy, có rất nhiều cách để chọn và nhà nghiên cứu sẽ phải quyết định cái gì là phù hợp nhất đối với nội dung nghiên cứu (chẳng hạn như FEV – một biến liên tục). Mặc dù vậy, liệu một lần đo FEV là đủ hay là người công nhân phải được theo dõi trong một vài ngày hay có thể là một năm? Kiến thức về biến phụ thuộc và về độ tin cậy của phép đo trên thực tế, và hơn nữa những hiểu biết về sinh lý học của các nguy cơ gây bệnh về đường hô hấp, sẽ giúp chúng ta đưa ra câu trả lời cho vấn đề này.

Biến độc lập là loại công nhân (làm việc ở khu vực trộn nguyên liệu và khu lò nung) – trong trường hợp này là một biến phân loại rất rõ ràng. Phần lớn các nghiên cứu định lượng thường có sự so sánh, có thể là giữa hai hay nhiều nhóm hoặc so sánh qua thời gian trên cùng một nhóm hay kết hợp nhiều so sánh. Trong trường hợp này chúng ta sẽ chọn so sánh FEV giữa hai nhóm công nhân tại một thời điểm.

Thông thường, giả thuyết khoa học được viết thành hai mệnh đề, *giả thuyết không* và *đối thuyết*. *Đối thuyết* là những gì mà nhà nghiên cứu thực sự tin hay mong đợi là sẽ đúng hoặc là hy vọng tìm ra kết quả là đúng khi nghiên cứu. *Giả thuyết không* thể hiện sự trung tính hoặc một kết quả đối lập. Trong ví dụ trên chúng ta có cặp giả thuyết là:

- Giả thuyết: (còn được gọi là H_0): FEV của hai nhóm công nhân đốt lò và công nhân làm tại khu trộn nguyên liệu là như nhau.
- Đối thuyết: (còn được gọi là H_1): FEV của hai nhóm công nhân khác nhau.

Đây còn gọi là kiểm định hai phía. Tuy nhiên, giả thuyết khoa học cũng có thể được đặt ra theo một cách khác, nếu nhà nghiên cứu tin chắc rằng người công nhân làm việc tại khu lò nung sẽ có nhiều nguy cơ bị mắc các bệnh về đường hô hấp hơn so với công nhân làm các công việc vận chuyển và trộn nguyên vật liệu:

- H_0 : FEV của nhóm công nhân làm tại khu trộn thì bằng hoặc tốt hơn so với công nhân tại khu lò nung.
- H_1 : FEV₁ của nhóm công nhân làm việc trộn nguyên liệu thì kém hơn so với công nhân làm việc tại khu lò nung.

Đây còn gọi là kiểm định một phía. Tuy nhiên, thông thường người ta hay dùng kiểm định hai phía hơn mặc dù đã nghĩ về “hướng” của phép so sánh. Kiểm định hai phía thường “an toàn hơn” và cho phép nhà nghiên cứu đưa ra báo cáo kết quả ý nghĩa thống kê theo cả hai hướng (kể cả khi thu được những kết quả không mong đợi).

Tất cả các nghiên cứu định lượng tập trung vào việc thu thập đủ thông tin để bác bỏ H_0 (mặc dù trên thực tế chúng ta làm nghiên cứu vì cho rằng H_1 là đúng). Kết luận cuối cùng của chúng ta thường sẽ được viết sao cho thể hiện rõ điều này, ví dụ:

Không đủ bằng chứng để bác bỏ H_0 , như vậy trong nghiên cứu này chúng ta kết luận là bệnh hô hấp không liên quan đến loại công việc của công nhân trong nhà máy gạch.

Hoặc:

Có đủ bằng chứng để bác bỏ H_0 , như vậy chúng ta kết luận là nghiên cứu này cho thấy vấn đề bệnh đường hô hấp có liên quan đến khu vực làm việc của công nhân trong nhà máy gạch. Những công nhân làm việc trong khu lò nung có nguy cơ suy giảm chức năng hô hấp nhiều hơn công nhân làm việc trong khu trộn nguyên liệu.



1.3. Các thiết kế nghiên cứu định lượng cơ bản

Có hàng loạt các thiết kế nghiên cứu cho phép làm giảm thiểu nguồn sai số ngẫu nhiên và hệ thống trong nghiên cứu. Hai loại thiết kế định lượng cơ bản là *thực nghiệm* và *quan sát* (không thực nghiệm). Thiết kế nghiên cứu thực nghiệm là dạng thiết kế được coi là lý tưởng nhất. Loại thiết kế này có ít nguy cơ xảy ra sự sai lệch trong kết quả. Tuy nhiên, do số lượng đối chứng cần thiết và quy trình kiểm soát nghiên cứu, loại thiết kế này thường tạo ra một bối cảnh nghiên cứu mang tính “nhân tạo” rõ rệt, ít phản ánh được thực tại. Điều này đôi khi làm cho dạng nghiên cứu này hoàn toàn không phù hợp, hoặc thậm chí vi phạm các quy định về đạo đức với các câu hỏi nghiên cứu có tính thực tế.

Có 3 đặc tính chính phân biệt thiết kế nghiên cứu thực nghiệm, đó là:

1. Có “can thiệp”, trong đó các đối tượng nghiên cứu được yêu cầu tham gia thực hiện các kiểm tra/hành vi/các hoạt động nào đó mà trong điều kiện thực tế cuộc sống họ chưa chắc đã phải làm.
2. Có một nhóm đối chứng, là nhóm đối tượng nghiên cứu không nhận được sự can thiệp nói trên.
3. Có sự phân bố ngẫu nhiên: các đối tượng được phân vào các nhóm khác nhau: nhóm can thiệp hay nhóm đối chứng. Hay nói cách khác, bất cứ một đối tượng nào cũng có một cơ hội bằng nhau để được chọn vào một trong hai nhóm. Ba đặc điểm riêng như tuổi, đặc điểm dân số, tiền sử về bệnh tương đương nhau ở các nhóm thì có thể giảm thiểu được các sai số do biến nhiễu.

Các nghiên cứu thực nghiệm là không khả thi, hoặc ít khả thi hơn những loại thiết kế khác và vì thế những loại thiết kế nghiên cứu “ít chặt chẽ hơn” thường được lựa chọn. “Ít” chặt chẽ hơn đồng nghĩa với việc khả năng dẫn tới sai số lớn hơn. Dưới đây liệt kê các loại thiết kế nghiên cứu được sắp xếp từ loại được kiểm soát tốt nhất (nghiên cứu thực nghiệm) cho tới kiểm soát kém nhất (nghiên cứu mô tả) thường được sử dụng trong các nghiên cứu y tế công cộng:

THỰC NGHIỆM

GIẢ THỰC NGHIỆM (còn gọi là “bán thực nghiệm” – quasi-experimental)

Nghiên cứu đánh giá sau can thiệp (Post test)

Nghiên cứu đánh giá trước/sau can thiệp (Pre-Post test)

QUAN SÁT

Thuần tập (Cohort)

Bệnh – Chứng (Case-Control)

Cắt ngang (Cross-sectional)

MÔ TẢ

Mô tả nhiều trường hợp (Case-series)

Mô tả trường hợp (Case-study)

Trên thực tế, khi thiết kế nghiên cứu có can thiệp nhưng lại không có sự phân bố ngẫu nhiên hoặc không có nhóm chứng thì được gọi là thiết kế nghiên cứu giả thực nghiệm (hoặc bán thực nghiệm).

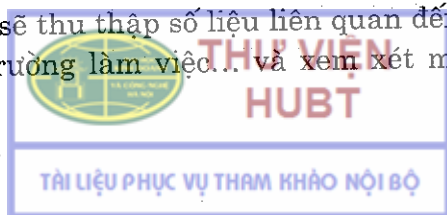
Tất nhiên, chất lượng nghiên cứu sẽ bị giảm khi thiếu đi các đặc điểm của thiết kế thực nghiệm, chủ yếu là nguy cơ xuất hiện sai số do sự không cân bằng giữa các đặc tính của các nhóm. Việc thiếu nhóm chứng có thể làm giảm khả năng phân tích các mối quan hệ nhân quả, vì chúng ta mất đi khả năng xác định các hiệu quả của can thiệp hay khẳng định những kết quả đó là vượt khỏi phạm vi những thay đổi tự nhiên có thể xảy ra. Tuy nhiên do giảm sự kiểm soát chặt chẽ trong thiết kế, thiết kế giả thực nghiệm khá linh hoạt và thường được sử dụng nhiều trong thực tế. Chúng ta cần lưu ý khi phiên giải kết quả của nghiên cứu giả thực nghiệm, đặc biệt khi chúng được dùng rất phổ biến trong các nghiên cứu về sức khỏe.

Tiếp theo trong danh sách phân loại chất lượng thiết kế là các nghiên cứu không thực nghiệm. Các thiết kế này không thực hiện các can thiệp hay phân bố ngẫu nhiên đối tượng nghiên cứu vào từng nhóm, chúng dựa trên cơ sở duy nhất là quan sát những gì đã hay sẽ xảy ra. Trong một số tình huống, một số thiết kế kiểu này cũng có thể có nhóm chứng nhưng không cho phép thực hiện các can thiệp. Những nghiên cứu thuộc dạng đó không có được sự kiểm soát chặt chẽ, thường có khuynh hướng dễ mắc các sai số.

Các thiết kế nghiên cứu quan sát gồm hai dạng cơ bản là mô tả và tương quan.

Nghiên cứu mô tả được thiết kế để mô tả tóm tắt các sự vật hiện tượng, và thông thường các mối quan hệ giữa những biến số trong dạng nghiên cứu này ít khi được nhấn mạnh. Ví dụ, người ta có thể thiết kế một nghiên cứu để xác định tỷ lệ người trong một cộng đồng sử dụng phương pháp châm cứu để chữa đau. Nghiên cứu mô tả thông thường được thiết kế để cung cấp những thông tin cơ bản và là dạng thiết kế dễ sinh ra sai số nhất (chủ yếu là sai số chọn và nhiễu).

Nghiên cứu tương quan đưa ra các mối liên quan giữa biến độc lập và biến phụ thuộc. Thông thường loại nghiên cứu này dễ sinh ra các giả thuyết. Ví dụ, *đặc điểm nghề nghiệp nào có ảnh hưởng đến việc hài lòng với nghề nghiệp?* Để làm điều này chúng ta sẽ thu thập số liệu liên quan đến nghề nghiệp, như số giờ làm việc, lương, môi trường làm việc... và xem xét mối liên hệ của chúng với



một thang điểm về sự hài lòng về nghề nghiệp. Chúng ta có thể không thu được một nhận định chính xác là liệu lương được bao nhiêu và môi trường làm việc như thế nào thì quyết định mức độ hài lòng về công việc nhưng nghiên cứu tương quan của chúng ta sẽ có thể xác định những đặc tính nghề nghiệp nào có thể liên quan đến sự hài lòng về nghề nghiệp và tạo tiền đề cho các nghiên cứu thực nghiệm về vấn đề này (chẳng hạn, liệu những người được chọn ngẫu nhiên vào nhóm nhận được can thiệp nâng cao kỹ năng làm việc nhóm có điểm hài lòng cao hơn những người làm nhóm công việc khác hay không?). Nghiên cứu thực nghiệm có can thiệp đó có thể cung cấp bằng chứng về nguyên nhân trực tiếp cho sự hài lòng về công việc hơn là nghiên cứu tương quan. Tuy nhiên, nghiên cứu thực nghiệm chỉ có thể tiến hành với một hay một số rất ít bối cảnh nơi làm việc, vì đòi hỏi phải có sự kiểm soát chặt chẽ, trong khi nghiên cứu tương quan – vì chỉ đơn thuần là một nghiên cứu quan sát – có thể xem xét rất nhiều yếu tố cùng một lúc.

Có một số dạng thiết kế nghiên cứu quan sát cụ thể trong từng loại nghiên cứu tương quan và nghiên cứu mô tả. Sau đây là các thiết kế nghiên cứu thông thường nhất:

- Nghiên cứu tương quan trong y tế công cộng bao gồm điều tra cắt ngang, nghiên cứu bệnh chứng và nghiên cứu thuần tập. Nghiên cứu cắt ngang điển hình thường được thực hiện dưới dạng một cuộc điều tra và đưa ra một bức tranh tại một thời điểm về một số vấn đề sức khỏe hay các yếu tố nguy cơ liên quan đến vấn đề đó. Tổng điều tra dân số là một ví dụ về nghiên cứu cắt ngang.
- Nghiên cứu thuần tập là một nghiên cứu trong đó các thành viên được theo dõi qua một thời gian và người ta đếm sự xuất hiện của một số sự kiện (thông thường là các trường hợp bệnh mới). Các đối tượng trong nghiên cứu thuần tập được chọn từ một bộ phận các cá thể có cùng một đặc điểm chung nào đó (ví dụ tất cả những người sống tại cùng một vùng địa lý tại thời điểm bắt đầu nghiên cứu, học cùng một trường, làm cùng một khu công nghiệp). Nghiên cứu thuần tập là nghiên cứu dọc, liên quan đến việc tiến hành các phép đo lường lặp đi lặp lại theo thời gian.
- Một thiết kế nghiên cứu bệnh chứng thu thập số liệu theo phương pháp hồi cứu, ngược với nghiên cứu thuần tập. Trong nghiên cứu bệnh chứng, chúng ta xác định các trường hợp bệnh trước tiên, rồi thu thập số liệu liên quan tới những đặc tính của đối tượng trong khoảng thời gian trước khi bệnh xảy ra. Chúng ta cũng làm tương tự với nhóm đối chứng, những người không bị bệnh nhưng cũng có những đặc điểm nhất định tương tự như các trường hợp bệnh. Vì vậy, có hai nhóm được chọn vào nghiên cứu. Thiết kế nghiên cứu này rất thông dụng khi cần nghiên cứu các sự kiện

hay các bệnh hiểm (trong khi nếu làm nghiên cứu thuần tập thì có thể chúng ta sẽ phải tiến hành trong nhiều năm mới có thể có được thậm chí chỉ một số nhỏ các sự kiện).

Đó là ba loại thiết kế nghiên cứu cơ bản thường được dùng trong nghiên cứu YTCC, tuy nhiên còn có rất nhiều các thiết kế nghiên cứu định lượng khác. Khái niệm quan trọng nhất mà chúng ta cần nhớ là với các thiết kế nghiên cứu khác nhau thì khả năng gặp phải sai số cũng sẽ khác nhau, và điều này cần được lưu ý tới khi phân tích.

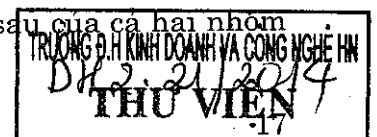
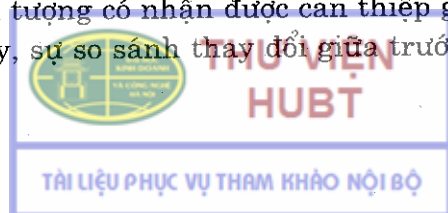
1.4. Các thành phần của thiết kế có ảnh hưởng tới việc phân tích

Với quan điểm của người phân tích số liệu, một số vấn đề thiết kế đóng vai trò quan trọng trong khâu phân tích. Dưới đây là một số vấn đề bạn cần xem xét tới khi tiến hành phân tích thống kê, bất kể việc phân tích đó đơn giản tới mức nào. Phần 1.4.4. sẽ có một bảng kiểm tóm tắt quy trình phân tích số liệu cần lưu ý.

1.4.1. Đơn vị quan sát

Một đặc điểm quan trọng thể hiện sự khác nhau giữa các loại thiết kế là sự khác biệt giữa đơn vị quan sát và các quan sát trong một nghiên cứu. Một đơn vị quan sát là một thành phần (ví dụ như người, động vật, cây, vùng địa lý, v.v.) được nghiên cứu. Nếu chúng ta chỉ thu thập số liệu một lần cho mỗi đơn vị quan sát thì số các quan sát sẽ bằng với số đơn vị quan sát. Tuy nhiên, trong các nghiên cứu có nhiều phép đo lường được tiến hành lặp lại trên cùng một đơn vị quan sát thì số các quan sát sẽ nhiều hơn số đơn vị quan sát. Trong phân tích số liệu, số các đơn vị quan sát là rất quan trọng và là nhân tố ảnh hưởng tới các phép tính toán. Một nghiên cứu nghe có thể rất ổn xét về phương diện lượng số liệu thu thập được khi có tất cả là 40 quan sát nhưng lại không ổn xét về mặt thống kê nếu chúng ta biết thêm rằng trong nghiên cứu đó người ta thu thập tới 20 quan sát trên mỗi đối tượng nghiên cứu, và vốn vẹn chỉ có tất cả là 2 đối tượng nghiên cứu. Trong thống kê thì nếu số đơn vị quan sát sẽ tốt hơn số quan sát sẽ tốt hơn.

Các thiết kế nghiên cứu thu thập số liệu nhiều lần trên cùng một người được gọi là nghiên cứu đo lường lặp lại hoặc thiết kế trên cùng một đối tượng (within-subject). Các thiết kế thu thập số liệu 1 lần trên mỗi đối tượng nhưng so sánh các đối tượng không liên quan với nhau được gọi là thiết kế giữa các đối tượng (between-subject). Một số loại thiết kế có thể có cả hai thành phần, ví dụ: một bộ câu hỏi được hỏi trước và sau khi tiến hành một can thiệp giáo dục, trong đó một nhóm đối tượng có nhận được can thiệp giáo dục còn nhóm kia thì không. Trong ví dụ này, sự so sánh thay đổi giữa trước và sau của cả hai nhóm



chính là so sánh *trên cùng đối tượng*, còn sự so sánh giữa nhóm không qua can thiệp và nhóm có nhận được can thiệp là so sánh *giữa các đối tượng*.

Các công thức thống kê cho phép phân tách phương sai của từng thành phần (giữa các đối tượng và trên cùng đối tượng). Vì thế, nếu chúng ta bỏ qua thực tế là một số quan sát của chúng ta bắt nguồn từ thiết kế *trên cùng đối tượng* thì các phép tính về sai số chuẩn và khoảng tin cậy sẽ không chính xác.

1.4.2. Phương pháp chọn mẫu

Một vấn đề thiết kế khác có ảnh hưởng tới việc tóm tắt và phân tích kết quả là các đối tượng được chọn vào nghiên cứu như thế nào? Một mẫu có thể được chọn từ một tập hợp (quần thể) theo nhiều cách khác nhau. Cách chọn lý tưởng nhất là chọn mẫu ngẫu nhiên đơn (random sample), trong đó phải xác định một danh sách các đối tượng phù hợp (khung mẫu) và sử dụng bảng số ngẫu nhiên để chọn một mẫu từ khung mẫu. Nếu được tiến hành đúng, mẫu chọn theo cách này thường sẽ đại diện cho quần thể và rất khách quan. Nhờ đó, người nghiên cứu không thể có cơ hội để quyết định đối tượng này thích hợp hơn đối tượng khác (chẳng hạn, vì họ dễ tiếp cận hơn, hay vì họ sẵn lòng trả lời câu hỏi phỏng vấn hơn). Việc này đảm bảo cho tất cả các đối tượng có một cơ hội được chọn vào mẫu nghiên cứu như nhau, và đó là điểm mấu chốt của phương pháp lấy mẫu ngẫu nhiên đơn.

Có nhiều biến thể của phương pháp lấy mẫu ngẫu nhiên đơn để giúp vượt qua những hạn chế về hậu cần khi tiến hành nghiên cứu. Mẫu ngẫu nhiên phân tầng được sử dụng để bảo đảm sự cân bằng về cỡ mẫu trong các phân nhóm mà chúng ta quan tâm. Ví dụ: Một mẫu ngẫu nhiên đơn gồm các ông bố hoặc bà mẹ sống độc thân có thể sẽ có nhiều phụ nữ hơn nam giới. Nếu như trong nghiên cứu này, chúng ta đặc biệt quan tâm đến việc có đủ cỡ mẫu để phân tích về các ông bố độc thân, chúng ta nên *phân tầng* khung mẫu của chúng ta thành hai nhóm nam và nữ và trong mỗi tầng (còn gọi nhóm) chúng ta chọn ngẫu nhiên số các đối tượng nam giới và nữ giới bằng nhau. Chọn như thế, chúng ta đã có một mẫu trong đó nam giới được chọn vượt tỷ lệ (over-representation) và sẽ đủ cỡ mẫu khi phân tích riêng nam giới. Khi cần tổng hợp số liệu cho cả hai giới, chúng ta sẽ giải quyết việc này bằng các thủ thuật thống kê (chẳng hạn như đặt *trọng số - weight*) vì nam và nữ được chọn với các xác suất khác nhau.

Thông thường, rất ít khi chúng ta có thể lấy mẫu theo phương pháp ngẫu nhiên đơn, đặc biệt là trong các điều tra quần thể. Người ta thường phải dùng một số cách chọn mẫu phức tạp hơn: chọn mẫu cụm (cluster sampling), chọn mẫu nhiều giai đoạn (multi-stage sampling). Chọn mẫu cụm thường được sử dụng để nghiên cứu các nhóm đối tượng đã phân bố theo từng "cụm" một cách tự nhiên (ví dụ như người dân sống thành từng làng, học sinh trong từng



trường, hay khu vực). Khi đó, một mẫu ngẫu nhiên các làng, trường... được chọn, và sau đó tất cả các thành viên tại từng địa bàn đó được chọn vào trong mẫu nghiên cứu, hoặc một số thành viên được chọn dựa trên nguyên tắc ngẫu nhiên.

Lấy mẫu phức tạp có phân cụm sẽ đưa thêm “sự phụ thuộc” vào trong bộ số liệu như trường hợp một đối tượng cung cấp nhiều quan sát đã mô tả trong phần 1.4.1. Vì thế, các phép tính phương sai sẽ phải được hiệu chỉnh trước khi có thể đưa ra sai số chuẩn và khoảng tin cậy một cách chính xác. Nếu chúng ta bỏ qua yếu tố thiết kế nghiên cứu có sử dụng các thủ thuật lấy mẫu phức tạp và coi như nghiên cứu sử dụng mẫu ngẫu nhiên đơn thì sai số chuẩn và khoảng tin cậy tính được sẽ không chính xác.

1.4.3. Các biến đầu ra

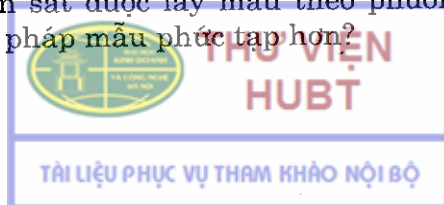
Như bạn sẽ thấy ở những phần tiếp theo, các loại biến số trong câu hỏi nghiên cứu sẽ ảnh hưởng rất lớn đến sự lựa chọn phương pháp phân tích thống kê. Mặc dù có rất nhiều loại biến khác nhau, cho mục đích phân tích, chúng ta hoàn toàn có thể đơn giản hóa chúng thành 2 loại là *biến liên tục* và *biến phân loại*. “Liên tục” ở đây bao gồm các biến liên tục thực sự và biến sử dụng thang đo khoảng (interval). Trong khi đó biến phân loại bao gồm biến nhị thức (dichotomous), định danh (nominal), và biến thứ bậc (ordinal). Biến số sử dụng là *liên tục* hay *phân loại* sẽ quyết định cách chúng ta tóm tắt số liệu (*trung bình* hay *tỷ lệ phần trăm*), hoặc cách chúng ta lựa chọn kiểm định thống kê. Trong các phần sau đây, bạn sẽ thấy rõ vai trò của từng loại biến mà bạn sẽ phân tích.

1.4.4. Bảng kiểm các thông tin để giúp bạn chuẩn bị cho phân tích thống kê

Thiết kế nghiên cứu sẽ quyết định phương pháp thu thập số liệu và rồi điều này sẽ ảnh hưởng tới các loại biến số dành cho khâu phân tích. Khi chuẩn bị tiến hành phân tích một bộ số liệu bạn nên bắt đầu bằng việc thu thập và trả lời đủ các thông tin về thiết kế nghiên cứu như sau:

Câu hỏi nghiên cứu là gì? Thông thường có nhiều hơn một câu hỏi nghiên cứu. Cố gắng chuyển câu hỏi nghiên cứu thành giả thuyết khoa học có thể kiểm định được (giả thuyết không và đối thuyết). Các bước sau đây sẽ giúp bạn làm điều đó:

1. Đó là thiết kế nghiên cứu thực nghiệm hay quan sát ?
2. Đơn vị quan sát là gì?
3. Các đơn vị quan sát được lấy mẫu theo phương pháp ngẫu nhiên đơn hay các phương pháp mẫu phức tạp hơn?



4. Các phép đo lường có được lặp lại (theo thời gian) hay có liên quan với nhau (chẳng hạn trên cùng một vị trí, hay một người) không?
5. Đầu ra của nghiên cứu là gì? thông thường có nhiều hơn một đầu ra
6. Biến đầu ra là biến liên tục hay biến phân loại?
7. Những nhóm nào được so sánh? có thể so sánh khác biệt của các nhóm đối tượng nghiên cứu khác nhau hoặc so sánh sự khác biệt của cùng một nhóm qua các mốc thời gian khác nhau.

Phân tích thống kê thích hợp cho một bộ số liệu phụ thuộc vào các câu trả lời cho những câu hỏi trên. Các bài học tiếp theo sẽ cung cấp các công cụ thống kê hoặc những cách tiếp cận để phân tích tính toán cho các tình huống khác nhau. Mục đích của bài trình này hướng dẫn bạn là nhận ra sự khác nhau, lựa chọn và áp dụng các kỹ thuật thống kê cơ bản phù hợp để phân giải các kết quả phân tích.

Bài số

2

QUẢN LÝ SỐ LIỆU

Mục tiêu

1. Hiểu được quá trình chuẩn bị một bộ số liệu nghiên cứu để nhập liệu.
2. Hiểu được lý do cơ bản, các nguyên tắc liên quan đến việc mã hoá số liệu và định nghĩa mã của một bộ số liệu.
3. Xác định và xây dựng một kế hoạch làm sạch số liệu.
4. Nhận biết được các chiến lược khác nhau cho việc đảm bảo tính toàn vẹn của bộ số liệu.

Trước khi phân tích số liệu, điều cần thiết là bạn mã số liệu đã thu thập dưới dạng số để phù hợp cho máy tính phân tích. Sau khi các số liệu được mã hoá, chúng ta sẽ nhập số liệu vào máy tính, tốt nhất là nhập trực tiếp vào một phần mềm thống kê. Trước khi việc phân tích có thể bắt đầu, số liệu cần phải được kiểm tra về tính chính xác và đầy đủ. Phần này liên quan với các vấn đề chuẩn bị và sàng lọc số liệu trước khi phân tích. Trên thực tế, phần lớn thời gian “phân tích” số liệu là dành cho việc chuẩn bị số liệu. Việc chuẩn bị số liệu bao gồm việc mã số liệu, nhập số liệu và kiểm tra thật cẩn thận bộ số liệu trước khi phân tích. Chuẩn bị số liệu là bước cơ bản để đảm bảo rằng phân tích của bạn là chính xác và đại diện cho số liệu bạn thu thập. Không nên đánh giá thấp tầm quan trọng của bước này, cũng như số lượng thời gian liên quan đến việc đạt được tính toàn vẹn của bộ số liệu.

2.1. Bộ số liệu mẫu

Phần này giới thiệu cho bạn khái niệm về quản lý số liệu, và giới thiệu một bộ số liệu trích ra từ Nghiên cứu Chấn thương trên toàn quốc năm 2001¹, bộ số

¹ Bộ số liệu sử dụng trong chương trình giảng dạy được rút ra từ nghiên cứu chấn thương trên toàn quốc do Đại học Y tế Công cộng phối hợp cùng 8 trường đại học và Viện Y tiến hành. Bộ số liệu dùng để giảng dạy ở đây đã được chỉnh sửa một phần so với bộ số liệu gốc với mục đích để phù hợp với nội dung của bài học.

liệu này sẽ được dùng trong suốt quyển sách và khoá học này. Chủ đề nghiên cứu là chấn thương giao thông ở Việt Nam và câu hỏi nghiên cứu là:

Trong số những người bị chấn thương giao thông năm 2001, những tác động nào của chấn thương ảnh hưởng đến chất lượng cuộc sống của họ?

Số liệu đã được thu thập năm 2001. Nghiên cứu này là một nghiên cứu mô tả cắt ngang trên một mẫu đại diện toàn quốc với quần thể nghiên cứu là toàn bộ người dân Việt Nam. Mẫu nghiên cứu dựa trên cách chọn mẫu ngẫu nhiên đơn và bộ câu hỏi tự điền dành cho bất kỳ trường hợp chấn thương nào trong một năm trước thời điểm nghiên cứu, trường hợp chấn thương là các nạn nhân bị chấn thương mà phải nghỉ học hoặc nghỉ việc ít nhất là một ngày. Nghiên cứu này thu thập số liệu mắc và tử vong của nhiều loại chấn thương khác nhau, tuy nhiên cuốn sách này chỉ sử dụng các số liệu liên quan đến các trường hợp chấn thương giao thông. Các biến dân số-xã hội (tuổi, giới, vùng, trình độ học vấn, nghề nghiệp) và các chi tiết về chấn thương giao thông (loại phương tiện giao thông có liên quan), hậu quả của chấn thương (vị trí chấn thương, số ngày nằm viện) cũng được thu thập. Các đối tượng nghiên cứu được hỏi và tự đánh giá về chất lượng cuộc sống trước và sau chấn thương. Lượng giá về chất lượng cuộc sống dựa trên thang điểm 100, bắt đầu từ 0 (chất lượng cuộc sống không thể thấp hơn) đến 100 (chất lượng cuộc sống không thể cao hơn được).

Các đối tượng nghiên cứu được gán một mã xác định và người phân tích và sử dụng bộ số liệu sẽ hoàn toàn không thể biết các thông tin cá nhân của đối tượng nghiên cứu.

2.2. Xử lý thông tin nghiên cứu cho phân tích định lượng

2.2.1. Xử lý và nhập số liệu

2.2.1.1. Mã hoá số liệu

Mục đích của việc mã hoá số liệu là chuyển đổi thông tin nghiên cứu đã thu thập thành dạng thích hợp cho việc phân tích trên máy tính. Thường thì bạn sẽ sử dụng một bộ câu hỏi hoặc biểu mẫu thu thập số liệu khác nhau để thu thập số liệu. Để đưa ra được những kết luận từ nghiên cứu của mình, bạn sẽ phải tóm tắt các kết quả của cuộc điều tra. Hầu hết các nghiên cứu đều liên quan đến một số lượng lớn các đối tượng tham gia, các thông tin từ bộ câu hỏi và các phiếu điều tra nên được nhập vào các phần mềm thống kê (như EpiInfo hay SPSS) để cho tất cả các số liệu đã thu thập được lưu trữ và xử lý một cách thuận tiện.

Phần mềm thống kê thường có dạng bảng tính và số liệu phải được nhập vào các bảng tính này. Một bảng tính là một định nghĩa tương đối về thực chất nó là

phần hiển thị trên màn hình với những đường kẻ trong đó. Các chữ số và các ký tự có thể được đánh máy trong từng ô. Khác với những phần mềm bảng tính thông dụng như EXCEL®, các phần mềm phân tích thống kê thường có những giới hạn chặt chẽ hơn trong việc nhập liệu vào những ô này. Trong hầu hết các tệp số liệu, thông tin cho từng đối tượng/quan sát được nhập trên một hàng của bảng tính. Các cột của bảng tính tương ứng với các câu hỏi trong bộ câu hỏi/công cụ thu thập số liệu. Một câu trả lời của một đối tượng được mã trên một ô của bảng tính (ví dụ một ô = một câu trả lời và tất cả các câu trả lời cho một câu hỏi nằm trên một hàng dọc dưới cột tương ứng). Các câu trả lời nên được mã hoá bằng số càng nhiều càng tốt và ta nên hạn chế việc dùng mã bằng các ký tự.

Ví dụ:

Mẫu phiếu trong bộ số liệu có dạng như sau:

ĐIỀU TRA CHẤN THƯƠNG GIAO THÔNG QUỐC GIA NĂM 2003

Mã hộ: _____

Vùng: _____

Ngày phỏng vấn: _____

Thông tin về người trả lời

1. Anh/chị bao nhiêu tuổi (tính tròn năm)? _____

2. Anh/chị là nam hay nữ (khoanh vào lựa chọn phù hợp)?

Nam *Nữ*

3. Anh/chị sống ở đâu? (Khoanh vào lựa chọn phù hợp)

Thành phố *Nông thôn*

4. Anh/chị đã học đến lớp mấy (Khoanh vào lựa chọn phù hợp)?

Không đi học *Tiểu học* *Dưới THCS* *Trên THCS*

Học nghề *Đại học* *Trên đại học* *Trẻ nhỏ*

5. Anh/chị làm nghề gì (Khoanh vào lựa chọn phù hợp)?

Làm ruộng *CBCNVC* *Buôn bán nhỏ* *Làm chủ*

Học sinh *Nghề thủ công* *Nghỉ hưu* *Mất sức*

Thất nghiệp *Trẻ nhỏ* *Khác*



Thông tin về chấn thương

6. Khi chấn thương xảy ra anh/chị là (Khoanh vào lựa chọn phù hợp):

Người lái xe Hành khách Người đi bộ Khác

7. Anh/chị đi bằng phương tiện gì khi chấn thương xảy ra (Khoanh vào lựa chọn phù hợp)?

Xe ô tô Xe máy Xe đạp Người đi bộ Khác

8. Anh/chị bị chấn thương nặng nhất ở đâu (Khoanh vào lựa chọn phù hợp)?

Đầu / cột sống Thân mình Tay / chân

9. Anh/chị có phải vào viện vì chấn thương này không? (Khoanh vào lựa chọn phù hợp)?

Có Không

10. Nếu có, anh/chị đã phải nằm viện bao nhiêu ngày? _____

Thông tin về chất lượng cuộc sống

11. Với thang điểm từ 0 (không thể thấp hơn) đến 100 (không thể cao hơn), anh/chị ước lượng chất lượng cuộc sống qua sức khỏe của anh/chị trước chấn thương là bao nhiêu?

12. Hiện nay anh/chị ước lượng chất lượng cuộc sống qua sức khỏe của anh/chị là bao nhiêu?

Bảng số liệu có dạng như sau:

h_id	Region	Age	gender	education	qol_bef
10111001	01	36	1	3	34
10121003	02	28	2	8	71
72012051	03	47	1	1	63
...

Lưu ý rằng trong khi tên cột có thể là bất kỳ tên gì bạn muốn, hầu hết các phần mềm thống kê quy định chặt chẽ về độ dài và dạng. Để phù hợp, hãy dùng dấu gạch dưới “_” vào tên biến dài hơn 8 ký tự, không để khoảng trống giữa các



từ (ví dụ 'h_id' thay cho 'household id'), và không bao giờ bắt đầu bằng một con số (cho dù ký tự số có thể dùng ở bất kỳ vị trí nào khác trong tên cột). Nếu bạn thiết kế bộ câu hỏi một cách đúng đắn, hầu hết các bản câu hỏi sẽ dễ dàng có thể được mã bằng chữ số.

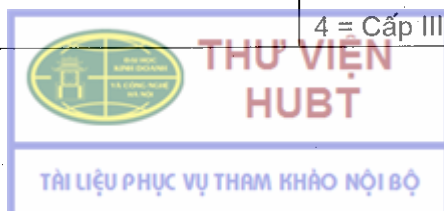
Mặc dù bạn chỉ được sử dụng tên cột không nhiều hơn 8 ký tự, hầu hết các phần mềm thống kê bao gồm cả SPSS đều cho phép bạn gán "nhãn" cho tên cột. Nhãn này có thể dài bao nhiêu tùy theo bạn muốn và sẽ xuất hiện thay vào tên cột trong phần kết quả của SPSS. Ví dụ bạn có thể đặt tên cột biểu thị cho tuổi là "age" nhưng bạn muốn nhớ rằng bạn đã nói về tuổi tính theo năm dương lịch hơn là theo tháng hoặc năm âm lịch. Bạn có thể gán một nhãn cho cột tên chẳng hạn "Tuổi dương lịch", nó sẽ xuất hiện trong phần kết quả khi bạn sử dụng biến này. Gán nhãn cho tên cột là một thói quen tốt để có thể dễ dàng xác định biến nào bạn đang sử dụng – đặc biệt nếu bạn có hai biến với tên tương tự hoặc đo lường những điều tương tự.

2.2.1.2. Bảng mã số liệu

Mã số cho từng câu trả lời nên được ghi lại chi tiết trong bảng mã số liệu.

Bảng mã số liệu

Biến	Mô tả	Giá trị mã và ý nghĩa
h_id	Mã hộ	Chạy từ 10111001–82436071
region	Địa điểm	1 = Tây Bắc 2 = Đông Bắc 3 = Lưu vực sông Hồng 4 = Bắc Trung bộ 5 = Ven biển miền Trung 6 = Cao Nguyên trung bộ 7 = Đông Nam bộ 8 = Lưu vực sông Mê Kông
ageround	Tuổi tính theo năm	0 – 65
sex	Giới tính	1 = Nam 2 = Nữ
u_r	Nơi cư trú	0 = Thành phố 1 = Nông thôn
educatio	Bậc học cao nhất	1 = Mù chữ 2 = Tiểu học 3 = Cấp II 4 = Cấp III



		<p>5 = Học nghề 6 = Đại học 7 = Sau đại học 8 = Trẻ nhỏ 99 = Không trả lời</p>
occupati	Nghề nghiệp	<p>1 = Làm ruộng/rẫy 2 = CBCNV 3 = Buôn bán nhỏ 4 = Buôn bán lớn (làm chủ) 5 = Sinh viên 6 = Trẻ dưới 6 tuổi 7 = Nghề thủ công 8 = Nghỉ hưu 9 = Không có khả năng làm việc 10 = Thất nghiệp 11 = Khác 99 = Không trả lời</p>
trantype	Loại phương tiện giao thông sử dụng khi xảy ra CT	<p>-1 = Không trả lời 1 = Ô tô 2 = Xe đạp 3 = Xe máy 4 = Người đi bộ 5 = Khác</p>
pedestrn	Nạn nhân là người đi bộ?	<p>0 = Không 1 = Có -1 = Không trả lời</p>
worst	Vị trí của chấn thương nặng nhất	<p>-1 = Không trả lời 1 = Đầu/cột sống 2 = Thân mình 3 = Tay chân</p>
hospital	Nạn nhân có nằm viện?	<p>-1 = Không trả lời 0 = Không 1 = Có</p>
q9	Số ngày nằm viện	<p>Giá trị từ 1-200 . = Không nằm viện</p>
qol_bef	Chất lượng cuộc sống trước chấn thương	<p>Giá trị từ 0 – 100</p>
qol_aft	Chất lượng cuộc sống sau chấn thương	<p>Giá trị từ 0 – 100</p>

Bao giờ cũng vậy, bạn sẽ nhận được những phiếu phỏng vấn mà có những câu hỏi không được trả lời. Trong tài liệu hướng dẫn nhập số liệu bạn nên quy định một mã cụ thể cho các trường hợp dữ liệu bị thiếu. Trong ví dụ ở trên, mã cho biến “hospital” khi không có câu trả lời là -1. Đó là cách tốt nhất phù hợp trong bảng mã hoá của bạn, nhưng bạn nên thường xuyên kiểm tra những giá trị đã được sử dụng để thay thế các thông tin bị mất. Ví dụ số liệu bị mất ở biến “occupati” được mã là 99. Một số câu hỏi thậm chí sẽ không được hỏi với một số đối tượng. Ví dụ, trong phiếu điều tra ở trên chúng ta chỉ hỏi “Anh/chị đã phải nằm viện bao nhiêu ngày?” với những trường hợp đối tượng phải nằm viện. Những người này có thể được mã là mất số liệu nhưng sẽ không thấy sự khác nhau giữa họ và những người có được hỏi nhưng không trả lời. Để giải quyết vấn đề này các đối tượng không được hỏi đã được mã là “.”, kí hiệu này chỉ ra rằng đối tượng này không phù hợp với câu hỏi.

Các vấn đề khác nảy sinh khi một đối tượng trả lời nhiều hơn một lựa chọn (ví dụ với câu hỏi “Anh/chị làm nghề gì” 3. Tiểu thương 4. Công chức 5. Sinh viên..., trong đó đối tượng có thể trả lời cả 3 và 5). Trong một vài trường hợp, có thể là quan trọng nếu chúng ta biết rằng đối tượng phỏng vấn là buôn bán nhỏ và chúng ta không cần biết họ cũng đang là sinh viên, vì thế trường hợp này nên chọn câu trả lời xác đáng nhất. Nếu thấy cả hai nghề nghiệp đều quan trọng, dùng giải pháp tạo hai cột trong bảng mã hoá để phù hợp với trường hợp này. Cột đầu tiên sẽ được gọi là EMPLOY1 và cột thứ hai sẽ là EMPLOY2 với các mã 1 và 2 tương ứng. Tất cả các trường hợp khác mà chỉ trả lời một lựa chọn thì mã vào EMPLOY1 và chọn mã không có số liệu trong EMPLOY2. Một giải pháp khác có thể sử dụng là chúng ta vẫn chỉ sử dụng một cột EMPLOY và mã thêm một giá trị nữa 3. Buôn bán nhỏ, 4. Công chức, 5. Sinh viên, và 6. Buôn bán nhỏ và sinh viên... Giải pháp đầu tiên linh hoạt hơn (và được ưa dùng hơn) nhưng đòi hỏi nhiều thời gian mã hoá hơn và có thể không hiệu quả.

2.2.2. Nhập số liệu

Một bộ câu hỏi hay một biểu mẫu số liệu được mã hoá dưới dạng số, chúng cần được nhập vào máy tính để chuẩn bị cho phân tích số liệu. Trừ khi các mẫu phiếu được thiết kế cho máy tính quét, phần nhập số liệu yêu cầu con người dùng tay và mắt để chuyển những thông tin, vì thế có thể xảy ra các lỗi đánh máy. Có nhiều cách nhập số liệu mà có thể hạn chế đến mức thấp nhất các lỗi đánh máy.

Cách tốt nhất là tạo một chương trình nhập số liệu sử dụng một phần mềm quản lý số liệu ví dụ như Microsoft Access, chương trình này có những chức năng để kiểm tra sự lặp lại của các mã hiệu hoặc các giá trị không có ý nghĩa. Tuy nhiên, sự phức tạp này cũng không đủ để tránh khỏi các lỗi đánh máy. Để



hạn chế mức thấp nhất sai lầm này, có một gợi ý đưa ra là tất cả số liệu nên được nhập hai lần, mỗi lần là một người khác nhau. Bất kỳ một sự khác nhau nào giữa hai bản số liệu cũng cần phải được lưu ý và phải được kiểm tra dựa trên bản số liệu gốc hoặc nếu có thể đối chiếu với đối tượng phỏng vấn. Nhập số liệu hai lần do hai người độc lập là lý tưởng, nó sẽ làm hạn chế tối đa các lỗi khi đọc và lỗi đánh máy. Ví dụ, chữ viết tay đôi khi không dễ đọc, một người có thể đọc là '3', nhưng người thứ hai có thể đọc là '5'. Sự khác nhau này sẽ được kiểm tra và mã số đúng sẽ được chọn. Nhập số liệu hai lần bởi cùng một người cũng có thể phát hiện ra lỗi đánh máy nhưng không phát hiện được lỗi đọc, cùng một người thì gần như vẫn đọc '3' là '3' trong cả hai lần.

Nhập số liệu hai lần tốn nhiều thời gian và đắt tiền. Rất khó thực hiện trên thực tế và đôi khi không cần thiết nếu chương trình nhập số liệu được viết và có nhiều chức năng kiểm tra quá trình nhập liệu. Dưới đây là các chiến lược nhập số liệu cho bạn lựa chọn, nhưng cần biết rằng khả năng mắc các lỗi đánh máy tăng lên khi danh sách này đi xuống. Vì thế, bạn sẽ cần phải có kế hoạch kiểm tra cẩn thận hơn trong phần làm sạch số liệu.

- i. Nhập toàn bộ số liệu hai lần bởi hai người riêng biệt.
- ii. Nhập toàn bộ số liệu hai lần do một người thực hiện.
- iii. Nhập toàn bộ số liệu một lần, sau đó chọn ngẫu nhiên đơn khoảng 20% bộ số liệu và nhập lần 2. Nếu những sự khác nhau là tối thiểu, dừng lại. Nếu không cần phải cân nhắc (ii).
- iv. Nhập toàn bộ số liệu 1 lần, chọn ngẫu nhiên đơn khoảng 20% bộ số liệu, kiểm tra lại bằng mắt. Nếu những sự khác nhau là tối thiểu, dừng lại. Nếu không cần phải cân nhắc (ii).
- v. Nhập toàn bộ số liệu một lần, không kiểm tra hai lần. Không có đề nghị gì.

Có rất nhiều các chương trình khác nhau để nhập số liệu, một vài chương trình rất phức tạp (ví dụ ORACLE®, SQL®), các chương trình khác thì ít phức tạp hơn (ví dụ dBase®, FoxPro®, Access®). Trong khóa học này EpiInfo sẽ được sử dụng để nhập số liệu như một hệ thống phần mềm miễn phí và dễ dàng sử dụng. Nếu bạn có phần mềm khác mà bạn thông thạo như Microsoft Access® hay FoxPro® thì cũng có thể sử dụng được. Hãy nhớ rằng lời cảnh báo về các lỗi đánh máy đưa ra ở trên và các bước làm sạch và quản lý số liệu liệt kê ở dưới sẽ áp dụng cho bất kỳ chương trình nhập số liệu nào bạn sử dụng.



XEM BỘ SỐ LIỆU TRONG SPSS

Mỗi lần bạn chuyển số liệu của bạn vào SPSS bạn sẽ thấy số liệu xuất hiện trong Data Window. Window có hai phần, Data View và Variable View.



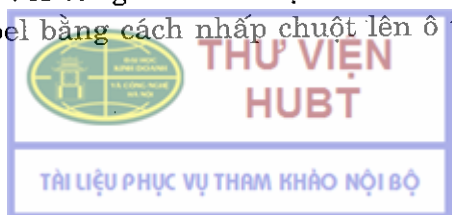
DATA WINDOW – DATA VIEW

	region	h_id	u_r	sex	occupati	q9	ageround	headp
1	1	10111001	1	1	1	7	0	
2	1	10111052	1	1	1	7	35	
3	1	10111055	1	1	3		37	
4	1	10111056	1	2	3	4	31	
5	1	10111061	1	1	1	1	41	
6	1	10111070	1	1	1	7	33	
7	1	10111090	1	1	1	1	43	
8	1	10112089	1	1	3		41	
9	1	10113019	1	2	1		53	
10	1	10113041	1	1	2		40	
11	1	10113089	1	2	5	1	13	
12	1	10114008	1	2	1	90	23	
13	1	10114039	1	1	1	3	26	
14	1	10114039	1	1	1	3	26	
15	1	10114063	1	1	4		25	
16	1	10114068	1	2	1	1	51	

DATA WINDOW – VARIABLE VIEW

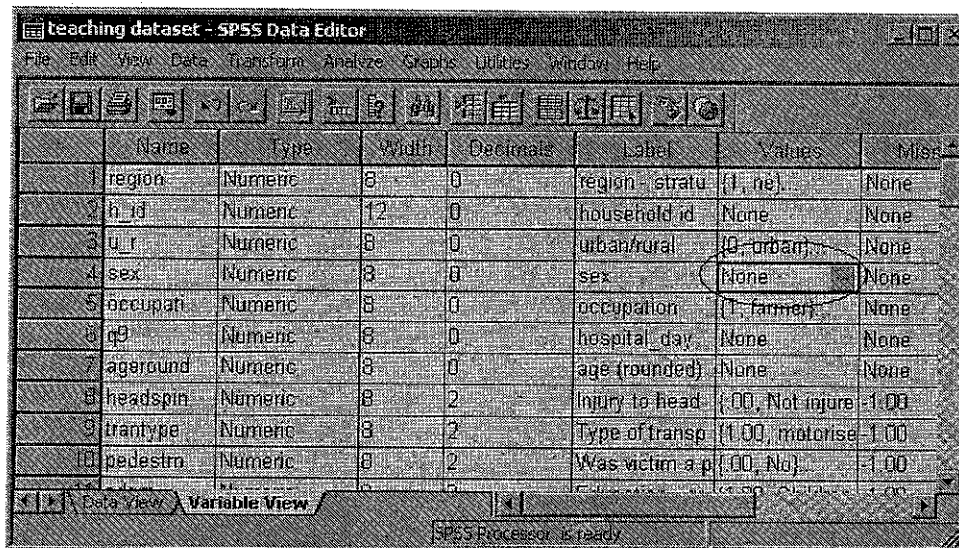
	Name	Type	Width	Decimals	Label	Values	Missing
1	region	Numeric	8	0	region - stratu	{1, ne}	None
2	h_id	Numeric	12	0	household id	None	None
3	u_r	Numeric	8	0	urban/rural	{0, urban}	None
4	sex	Numeric	8	0	sex	{1, male}	None
5	occupati	Numeric	8	0	occupation	{1, farmer}	None
6	q9	Numeric	8	0	hospital_day	None	None
7	ageround	Numeric	8	0	age (rounded)	None	None
8	headspin	Numeric	8	2	Injury to head	{1.00, Not injure}	-1.00
9	trantype	Numeric	8	2	Type of transp	{1.00, motorise}	-1.00
10	pedestrm	Numeric	8	2	Was victim a p	{0.00, No}	-1.00
11	edgrp	Numeric	8	2	Education - gr	{1.00, Children}	-1.00
12	qol_bef	Numeric	8	2	General quality	None	-1.00
13	qol_aft	Numeric	8	2	general quality	None	-1.00
14	diff	Numeric	8	2	Difference in q	None	None
15	qolbeig	Numeric	8	2	Quality of life b	{0.00, Suboptim	None
16	qolaftg	Numeric	8	2	Quality of life a	{0.00, Suboptim	None
17	hospital	Numeric	8	2	Hospitalised d	{1.00, No}	None

Để gán nhãn cho một biến, mở cửa sổ số liệu trong SPSS. Chọn Variableview ở góc dưới trái màn hình. Bạn sẽ thấy toàn bộ các tên cột (biến) từ trên xuống ở phía trái. Với từng tên biến bạn có thể đưa bất kỳ nhãn nào bạn muốn dùng vào cột Label bằng cách nhấp chuột lên ô thích hợp và nhập nhãn vào ô đó.

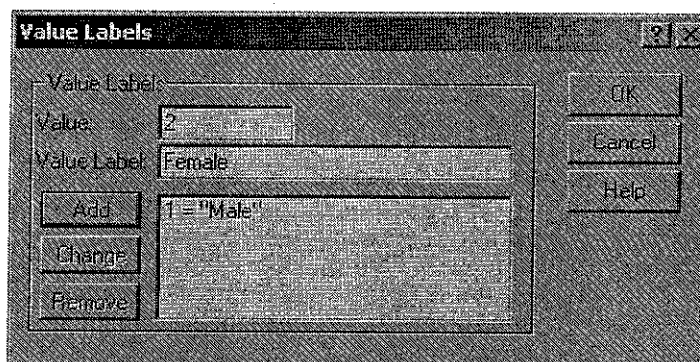


Bạn cũng nên gắn nhãn cho các mã số liệu, việc làm này có nhiều lợi ích, nó có tác dụng tạo ra một bảng mã điện tử của bộ số liệu. Để thêm các nhãn vào mã số liệu bạn làm như sau:

1. Vào Data Window – Variable View



2. Chọn biến bạn muốn gắn nhãn vào mã số liệu, chuyển chuột đến cột Values. Nhấp chuột lên ô này và một hộp nhỏ màu xám sẽ xuất hiện ở góc ô này. Nhấp chuột lên hộp đó và màn hình hiển thị một hộp nhỏ tương tự như dưới đây.



3. Để gắn thêm nhãn cho từng giá trị, nhập giá trị vào hộp Value, sau đó đưa nhãn bạn muốn cho vào trong hộp Value Label. Nhấp chuột vào Add. Giá trị với nhãn được gắn sẽ chuyển xuống hộp ở dưới và hai hộp ở trên sẽ trống cho phép bạn nhập các giá trị và nhãn khác.

4. Sau khi bạn đã đưa toàn bộ các nhãn xuống được hộp ở dưới (nên nhớ là bao gồm cả các mã cho số liệu bị mất), nhấp chuột vào OK. Các nhãn của bạn đã được gắn xong.

2.2.3. Làm sạch số liệu

Đưa số liệu từ phiếu phỏng vấn vào phần mềm máy tính có thể là một quá trình tốn thời gian và buồn tẻ, dễ có lỗi của con người ở nhiều điểm trong suốt quá trình này. Có rất nhiều khả năng chúng ta sẽ mắc lỗi trong quá trình điền phiếu, mã số liệu và nhập số liệu vào máy tính.

Có ba lỗi chính là:

- i. Các lỗi về mã số liệu.
- ii. Các lỗi về nhập số liệu.
- iii. Các lỗi về tính nhất quán (chắc chắn) của số liệu.

Để đảm bảo rằng bộ số liệu bạn đang xử lý là chính xác, bạn cần phải hạn chế đến mức tối đa những sai sót trong quá trình chuẩn bị và nhập số liệu.

Bước đầu tiên để kiểm tra số liệu của bạn là bạn liệt kê toàn bộ các giá trị của tất cả các biến trong bộ số liệu (bảng tần số). Việc làm này giúp bạn có thể quan sát nếu có bất kỳ lỗi nào về mã số liệu (ví dụ như mã là 5 trong biến giới tính ở ví dụ trên trong khi chỉ có mã 1 và 2 là hợp lý theo như bộ mã số liệu). Các giá trị khác thường có thể được liệt kê với những mã hiệu tương ứng và có thể được kiểm tra lại với phiếu gốc. Có thể các giá trị đáng nghi này là đúng với số liệu gốc thu thập được (ví dụ một chỉ số áp lực động mạch cao khác thường), bởi thế thay vì số liệu đáng ngờ bạn đã xác định được một giá trị bất thường (outlier) trong bộ số liệu của bạn. Các giá trị bất thường có thể được so sánh với phiếu gốc để kiểm tra xem chúng có chính xác không, nếu đúng chúng cũng phải được đưa vào trong bất kỳ một phân tích nào mặc dù giá trị đó có vẻ kỳ quặc.

Các lỗi mã hoá số liệu (coding errors) liên quan đến việc sai mã của đối tượng. Trùng mã xác định đối tượng là một lỗi phổ biến. Các lỗi khác cần phải xem xét bao gồm cả các lỗi mã số liệu. Ví dụ đối tượng trả lời “rất đồng ý” cho câu hỏi 1 (mã 1) nhưng lại được mã là “đồng ý” (mã 2) trong bảng số liệu. Cách tốt nhất để chắc chắn rằng không có lỗi mã hoá số liệu là đọc và sửa tệp số liệu từ phiếu gốc. Tuy nhiên, đọc và sửa chỉ thực hiện được với tệp số liệu nhỏ. Với những tệp số liệu lớn, các lỗi mã số liệu thường được kiểm tra bằng cách nhập lại toàn bộ hoặc chọn ngẫu nhiên từ 10–20% trường hợp của tệp số liệu khác và so sánh các kết quả giữa bản số liệu gốc và bản số liệu nhập để kiểm tra. Việc này được gọi là nhập kiểm tra (verification entry). Đó là một thành phần quan trọng để đảm bảo tính chân thực của số liệu cho dù nó có thể là một việc làm tốn kém. Không có việc nhập số liệu nào là hoàn hảo, nhưng nếu có ít hơn 1 trong 1000 bản ghi thông tin là không chính xác, thì cũng có thể cho rằng tác động của sai số nhập liệu lên các phân tích là tối thiểu.



Lưu ý: Bạn nên thường xuyên ghi chép lại những quyết định mã hoá số liệu của bạn. Chúng ta sẽ rất dễ quên các nguyên tắc đã dùng để mã, và trong một bộ số liệu việc thống nhất mã hoá theo một nguyên tắc là rất quan trọng. Nếu bạn không ghi chép lại các mã đã dùng thì sẽ có rất nhiều khả năng mắc lỗi hoặc mâu thuẫn trong khâu mã hoá số liệu.

Lỗi nhập số liệu (data entry errors) có thể xảy ra khi một mã bị đọc sai khi nhập số liệu vào máy tính (đọc bản viết tay là 5 khi trên bản viết là 3) hoặc lỗi đánh máy (đọc mã là 6 nhưng đánh máy là 5). Nhập kiểm tra được dùng để chữa những lỗi nhập liệu. Lý tưởng là có hai người đọc lập nhập số liệu hai lần riêng rẽ.

Khi bộ số liệu được nhập và “làm sạch” những lỗi đánh máy, còn có một mức độ làm sạch số liệu cao hơn để cân nhắc – đó là kiểm tra tính nhất quán (consistency checking). Điều này có nghĩa là những câu trả lời không nhất quán cần được xác định và kiểm tra. Một ví dụ về câu trả lời không nhất quán là với phiếu được mã là nam nhưng lại trả lời là “Có” cho câu hỏi “đã bao giờ mang thai chưa?”. Một ví dụ khác là ngày tử vong lại trước ngày sinh hoặc trả lời THCS như là bậc học cao nhất của trình độ học vấn nhưng sau đó lại mô tả về quá trình học đại học mà họ đã hoàn thành. Tính không nhất quán có thể là do các lỗi mã hoá số liệu hoặc đánh máy mà đã không bị phát hiện trong hai lần kiểm tra (nếu điều này xảy ra), hoặc đối tượng phỏng vấn trên thực tế đã đưa ra những câu trả lời không nhất quán. Nguyên nhân cuối cùng này yêu cầu phải liên lạc với đối tượng phỏng vấn để xác định lại nhưng điều này thường là không thể thực hiện được.

2.3. Các ví dụ về làm sạch số liệu

Trong các bài 3 và bài 4 bạn sẽ được giới thiệu về kế hoạch phân tích số liệu. Tuy nhiên, một kế hoạch có tính quan trọng trong việc quản lý số liệu là việc chuẩn bị số liệu sẵn sàng cho phân tích. Kế hoạch làm sạch số liệu của bạn phải không phức tạp. Bạn có thể sử dụng bản kế hoạch sau cho làm sạch số liệu như là một bảng kiểm, phần này sẽ đưa bạn đến việc làm thế nào để kiểm tra trong SPSS:

Kế hoạch làm sạch số liệu

1. Xác định các số xác định đối tượng (ID) trùng nhau.
2. Kiểm tra các giá trị bất thường (outliers) của tất cả các biến liên tục (ngoại trừ ngày tháng).
3. Kiểm tra các mã không phù hợp ở tất cả các biến phân loại (ngoại trừ biến ID).
4. Kiểm tra ngày tháng.

5. Kiểm tra số giá trị bị mất cho từng biến.

6. Định rõ những câu trả lời không nhất quán, bạn sẽ kiểm tra về:

Tuổi bằng 0

Tuổi sai khác khi lấy ngày sinh trừ đi ngày phỏng vấn.

Đối tượng có nằm viện khi không có chấn thương.

Không có sự phù hợp giữa trình độ học vấn và nghề.

Người đi bộ được phân loại là người lái xe.

Trẻ nhỏ lại đi học.

Trẻ em với bậc học cao hơn lứa tuổi.

Liệt kê những cá nhân có câu trả lời không nhất quán.

Lưu ý rằng những sự không nhất quán mà bạn chọn kiểm tra sẽ khác nhau giữa các bộ số liệu khác nhau, danh sách ở trên chỉ là một ví dụ. Bạn là người duy nhất có thể xác định phạm vi kiểm tra phù hợp cho bộ số liệu của bạn.

7. Nếu có thể hãy đối chiếu tất cả các lỗi với phiếu gốc, bạn sẽ xác định được các lỗi mã hoá hoặc nhập số liệu mà có thể chỉnh sửa được.

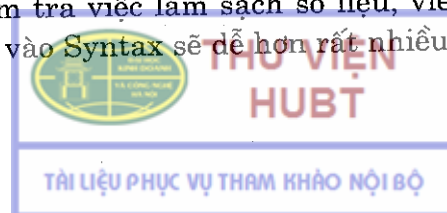
8. Nếu không có lỗi khi mã hoá cũng như nhập số liệu và các đối tượng phỏng vấn thực sự đã đưa ra các câu trả lời không nhất quán bạn nên liên lạc với họ để xác định lại thông tin. Tuy nhiên điều này thường là không thể thực hiện được.

2.3.1. Sử dụng SPSS để làm sạch số liệu

2.3.1.1. Thực đơn hay Syntax?

SPSS là một phần mềm rất thuận tiện cho người sử dụng. Có hai cách để sử dụng phần mềm này: cách thứ nhất là dùng thực đơn có trong Data Window, cách thứ hai là viết cú pháp/câu lệnh (syntax). Bài 3 và 4 sẽ cho bạn thấy làm thế nào để SPSS đưa ra những thống kê mô tả và suy luận từ thực đơn. Hầu hết các lệnh thống kê mô tả và suy luận đều ở thực đơn Analyse.

Khi bạn sử dụng thực đơn lệnh, SPSS đã đổi những chỉ dẫn bạn đưa ra từ thực đơn thành dạng ngôn ngữ của SPSS và được gọi là Syntax. Cách thứ hai để ra lệnh cho SPSS là viết một cách chính xác những gì bạn muốn làm bằng ngôn ngữ của chương trình này. Nó được viết trực tiếp vào Syntax Window trong SPSS, Syntax có thể được mở bằng lệnh File/New/Syntax từ thực đơn trong Data Window. Để kiểm tra việc làm sạch số liệu, viết chính xác những gì bạn muốn SPSS thực hiện vào Syntax sẽ dễ hơn rất nhiều việc sử dụng lệnh từ thực

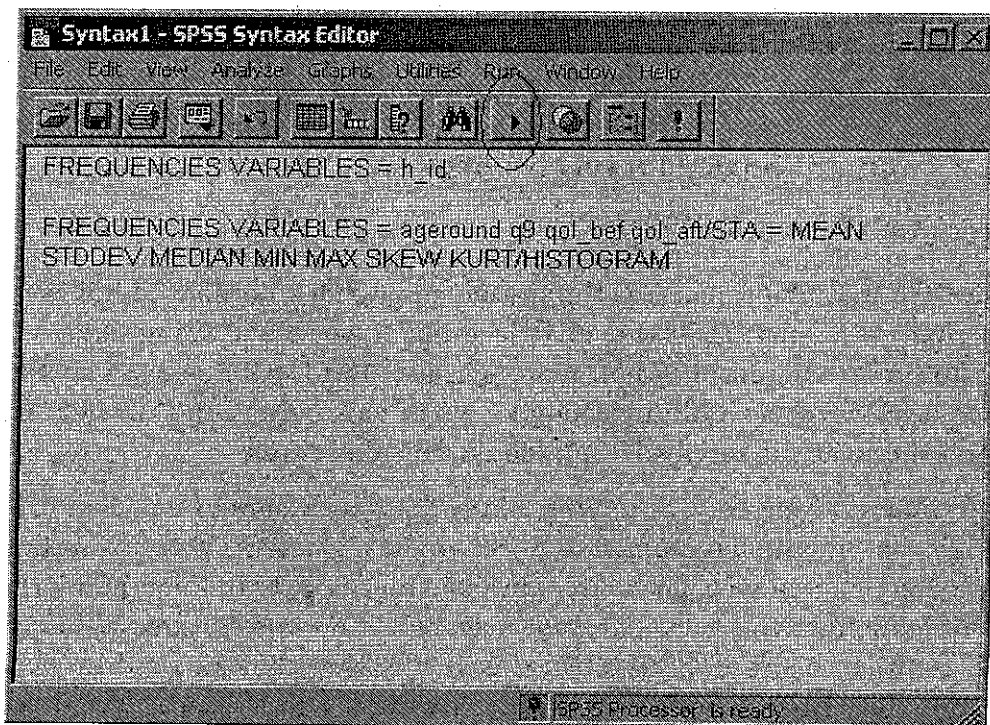


đơn. Bài này sẽ cung cấp cho bạn những *cú pháp* (syntax) mà bạn cần thực hiện cho từng kiểm tra và bạn có thể viết trực tiếp lên Syntax Window. Chép lại cú pháp thật cẩn thận (bao gồm cả khoảng trống, nét vạch chéo, dấu chấm) vì nó cần được viết một cách cực kỳ chính xác.

Ngoài việc nhanh chóng, lợi ích chính khác của việc sử dụng cú pháp là bạn có thể lưu lại tất cả những cú pháp bạn đã viết vào một tệp và bạn có thể sử dụng lại tệp này và thực hiện y hệt quá trình làm sạch số liệu và các phân tích như đã thực hiện trước đây. Điều này rất quan trọng trong trường hợp bạn cần kiểm tra những gì bạn đã phân tích đặc biệt với những thông tin mới được phát hiện. Bạn cũng có thể dùng lại các cú pháp này để phân tích một bộ số liệu mới sau khi đã chỉnh sửa cho phù hợp, điều này sẽ tiết kiệm rất nhiều thời gian cho bạn.

Ví dụ về dạng Syntax Window được đưa ra dưới đây. Bạn đánh máy những câu lệnh của bạn, bôi đen câu lệnh bạn muốn thực hiện và ấn lên hình mũi tên ở trên thanh công cụ. Kết quả sẽ xuất hiện trong Output Window như hình dưới đây.

SYNTAX WINDOW



OUTPUT WINDOW

The screenshot shows the SPSS Output Window with the following data:

Statistics

education		
N	Valid	4731
	Missing	20

education

		Frequents	Percent	Valid Percent	Cumulative Percent
valid	illiter	88	2.1	3.9	3.9
	primary	424	24.2	24.9	28.8
	lowerse	624	30.3	38.7	67.5
	upperse	355	20.8	49.9	88.4
	vocallon	78	4.7	4.5	92.9
	uneduc	7	1.5	1.5	94.4

Mặc dù bài 3 và 4 sẽ trình bày cho bạn cách thực hiện một phân tích sử dụng thực đơn, nhưng bạn có thể thấy rằng sử dụng cú pháp lệnh sẽ cho phép bạn thực hiện nhanh hơn, hoặc bạn có thể ghi lại những gì bạn đã làm sau khi sử dụng thực đơn. SPSS cho phép bạn dùng thực đơn dọc thực hiện các phân tích sau đó chuyển lệnh thực hiện đó sang Syntax bằng lệnh Paste. Nên nhớ, bạn cần nhấn nút Paste thay vì nhấn nút OK. Các cú pháp bạn tạo ra khi dùng thực đơn dọc sẽ xuất hiện trong Syntax Window. Kết quả sẽ không xuất hiện cho đến khi bạn chạy cú pháp này bằng cách bôi đen nó và nhấn lên mũi tên đen nằm phía dưới menu Run. Quá trình này được mô tả ở dưới trong phần “Kiểm tra sự giống nhau của số xác định đối tượng”.

2.3.1.2. Kiểm tra việc làm sạch số liệu trong SPSS

Kiểm tra sự giống nhau của số xác định đối tượng (ID):

Bạn cần kiểm tra sự giống nhau của số xác định đối tượng. Mỗi đối tượng khác nhau nên có một số xác định duy nhất để có thể xác định từng đối tượng. Nếu số xác định đối tượng bị trùng nhau thì bạn cần phải kiểm tra xem có phải một người đã bị nhập số liệu hai lần hay không, nếu có thì một bản ghi sẽ phải bị loại bỏ. Nếu không thì hệ thống đánh số của bạn đã sai và thực sự đây là bản ghi của hai người riêng biệt.

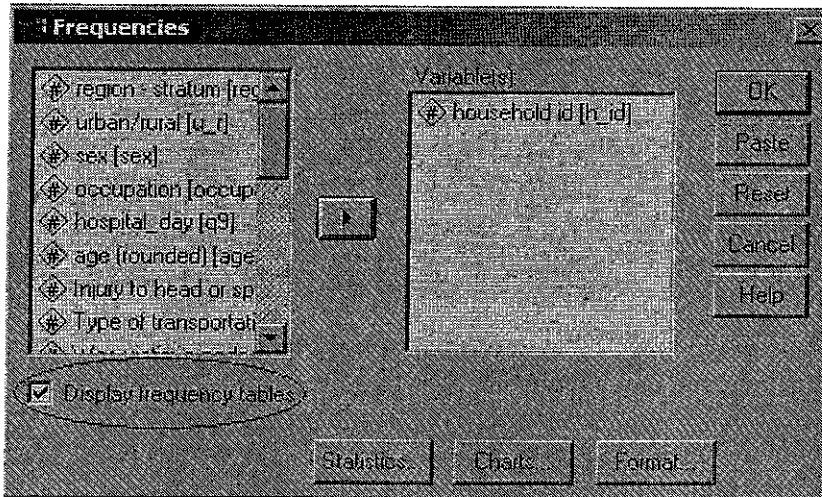
Để kiểm tra sự giống nhau của số xác định đối tượng, cách dễ dàng nhất là xem tần số của số xác định. Tần số của tất cả các số xác định phải là 1, nếu không số xác định đã bị trùng nhau và bạn nên quay lại kiểm tra phiếu phỏng vấn.

Sử dụng thực đơn

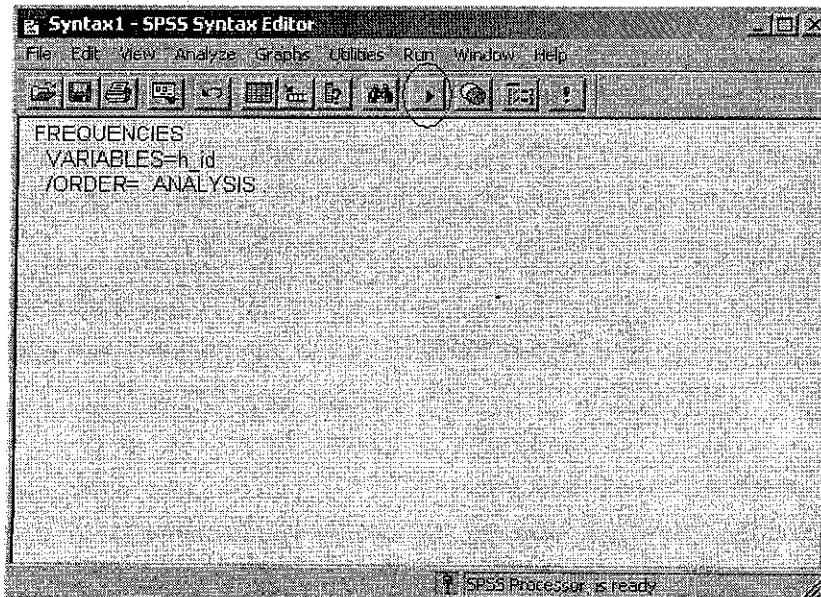
1. Từ thực đơn dọc chọn **Analyse** → **Descriptive Statistics** → **Frequencies**. Một hộp thoại sẽ xuất hiện như hình dưới đây.

2. Chọn biến bạn cần, trong trường hợp này là **h_id** (mã hộ gia đình), từ danh sách biến ở bên trái và chuyển vào trong hộp **Variable(s)** bên phải bằng cách nhấp chuột lên mũi tên.

3. Đảm bảo rằng **Display frequency tables** đã được chọn.



4. Nếu bạn nhấp chuột lên **OK** kết quả sẽ xuất hiện trong **Output Window** nhưng bạn sẽ không có bản ghi những gì bạn đã làm, vì thế nhấp chuột lên **Paste**. Cú pháp của bạn sẽ xuất hiện trong **Syntax Window** như hình dưới đây:



5. Bôi đen cú pháp này và nhấp chuột lên mũi tên ở thanh công cụ để thực hiện cú pháp. Kết quả sẽ xuất hiện trong các cửa sổ riêng rẽ – Output Window. Bạn có thể sửa lại câu lệnh này và thay h_id bằng bất kỳ tên biến nào biểu thị bằng số xác định và trực tiếp viết trên Syntax Window.

KẾT QUẢ

Đây là kết quả của cú pháp trên. Như bạn thấy có một số con số bị trùng và cần phải được kiểm tra lại.

Statistics		
household id		
N	Valid	1806
	Missing	0

		household id			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	10111001	1	.1	.1	.1
	10111011	1	.1	.1	.1
	10111052	1	.1	.1	.2
	10111056	2	.1	.1	.3
	10111061	1	.1	.1	.3
	10111070	1	.1	.1	.4
	10111090	1	.1	.1	.4
	10112019	1	.1	.1	.5
	10112089	1	.1	.1	.6
	10113019	1	.1	.1	.6
	10113041	1	.1	.1	.7
	10113089	1	.1	.1	.7
	10114008	1	.1	.1	.8
	10114039	2	.1	.1	.9
	10114063	1	.1	.1	.9
	10114068	1	.1	.1	1.0
	10114070	1	.1	.1	1.1

Kiểm tra những mã không phù hợp cho các biến phân loại

Bạn muốn đảm bảo rằng trong bộ số liệu của bạn không có giá trị không hợp lệ nào được nhập, ví dụ mã 5 cho biến giới tính khi mã hợp lệ chỉ là 1 hoặc 2. Điều này sẽ thực sự quan trọng khi bạn so sánh câu trả lời giữa hai nhóm Nam và Nữ, SPSS sẽ cố gắng và so sánh câu trả lời giữa 3 nhóm, Nam, Nữ và 5! Vì thế kết quả của bạn sẽ không đúng.



Để kiểm tra những giá trị (mã) không phù hợp của các biến phân loại, bạn có thể dùng lệnh Frequencies như đã nói ở trên để liệt kê ra các giá trị của biến đó. Nhìn vào bảng liệt kê, bạn sẽ biết được giá trị nào là không hợp lệ cho từng biến.

Bạn cũng có thể dùng syntax như sau để kiểm tra những mã không phù hợp cho các biến vùng, thành thị hay nông thôn, giới, nghề nghiệp, trình độ học vấn... Bạn nhập các tên biến tên một cách chính xác (tên biến trong cơ sở dữ liệu chứ không phải là nhãn của biến) mà bạn cần tính vào ô. Nếu có nhiều biến thì các biến cách nhau một dấu cách.

SYNTAX

FREQUENCIES VARIABLES = region u_r sex occupati educatio trantype q41_e worst hospital.

Cú pháp này yêu cầu các tần số (số tuyệt đối và tỷ lệ) của các biến phân loại. Bạn có thể thay tên biến (trong cú pháp trên tên biến là những ký tự nhỏ) bằng các tên biến khác trong bộ số liệu của bạn.

KẾT QUẢ:

Frequencies

Frequency Table

Region – stratum

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ne	200	11.6	11.6	11.6
	nw	56	3.3	3.3	14.9
	rr	200	11.6	11.6	26.5
	nc	259	15.0	15.0	41.5
	cc	152	8.8	8.8	50.4
	ch	154	8.9	8.9	59.3
	se	293	17.0	17.0	76.4
	mr	407	23.6	23.6	100.0
	Total	1721	100.0	100.0	

Urban/rural

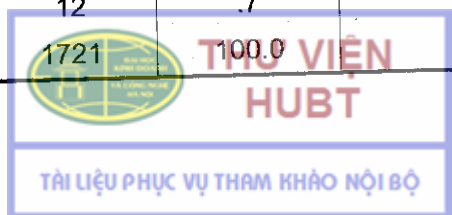
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid urban	539	31.3	31.3	31.3
rural	1182	68.7	68.7	100.0
Total	1721	100.0	100.0	

Sex

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid male	1107	64.3	64.3	64.3
female	614	35.7	35.7	100.0
Total	1721	100.0	100.0	

Occupation

	Freque	Percent	Valid Percent	Cumulative Percent
Valid farmer	546	31.7	31.9	31.9
gov. off	217	12.6	12.7	44.6
petty tr	84	4.9	4.9	49.6
employer	39	2.3	2.3	51.8
studying	309	18.0	18.1	69.9
children	129	7.5	7.5	77.5
handicra	190	11.0	11.1	88.6
retired	69	4.0	4.0	92.6
poor hea	22	1.3	1.3	93.9
unemploy	66	3.8	3.9	97.8
other	38	2.2	2.2	100.0
Total	1709	99.3	100.0	
Missing System	12	.7		
Total	1721	100.0		



Education

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	illitera	66	3.8	3.9	3.9
	primary	424	24.6	24.9	28.8
	lower se	624	36.3	36.7	65.5
	upper se	355	20.6	20.9	86.4
	vocation	76	4.4	4.5	90.8
	undergra	79	4.6	4.6	95.5
	graduate	3	.2	.2	95.6
	children	74	4.3	4.4	100.0
	Total	1701	98.8	100.0	
System	20	1.2			
Total		1721	100.0		

Type of transportation in which victim travelling

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	motorised vehicle	68	4.0	4.5	4.5
	Bicycle	332	19.3	21.8	26.2
	Motorised bike	889	51.7	58.3	84.5
	Pedestrian	168	9.8	11.0	95.5
	Other	68	4.0	4.5	100.0
	Total	1525	88.6	100.0	
Missing	-1.00	196	11.4		
Total		1721	100.0		

Most severe injury

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	head/spine	624	36.3	37.6	37.6
	torso	379	22.0	22.8	60.5
	limbs	656	38.1	39.5	100.0
	Total	1659	96.4	100.0	
Missing	-1.00	62	3.6		
Total		1721	100.0		

Hospitalised due to injury

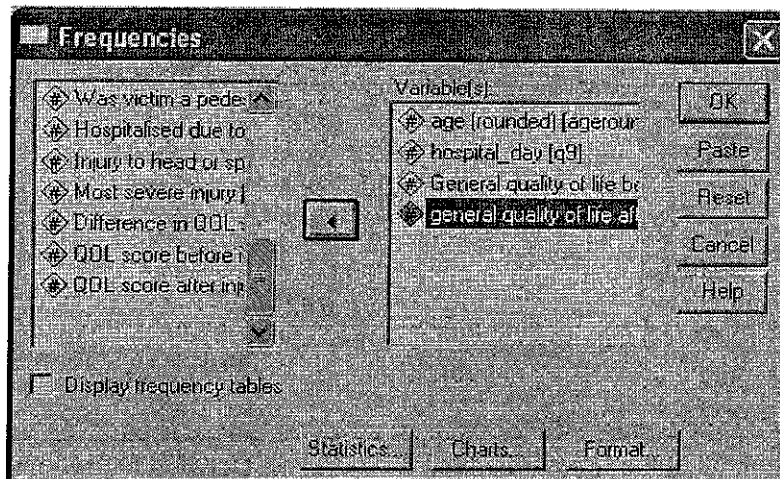
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	911	52.9	52.9	52.9
	Yes	810	47.1	47.1	100.0
Total		1721	100.0	100.0	

Kiểm tra các giá trị bất thường (outlier) của biến liên tục

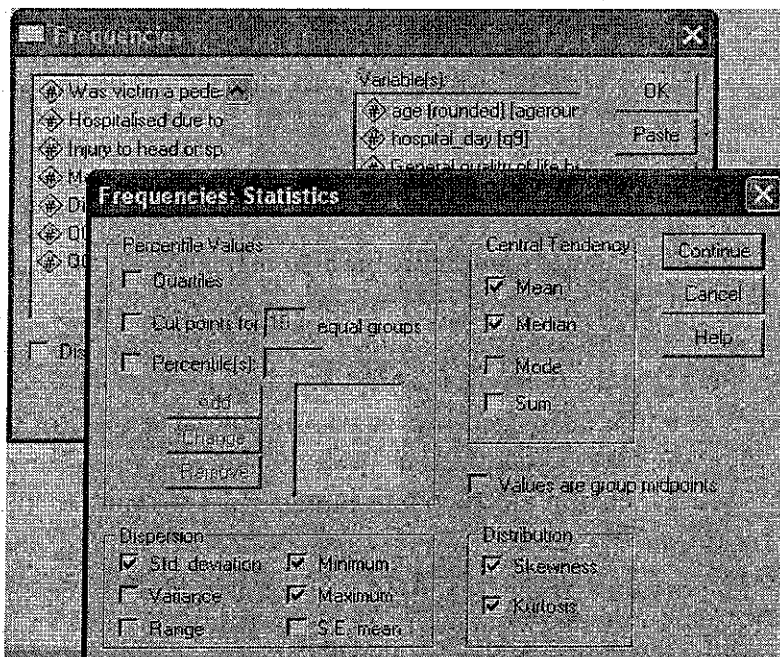
Kiểm tra các giá trị nằm ngoài của biến liên tục là một trong những kiểm tra quan trọng nhất bạn phải làm. Bạn muốn xác định bất kỳ một giá trị tột cùng nào (cả thấp và cao) trong bộ số liệu của mình. Hầu hết các kiểm định thống kê được mô tả trong bài 3 và 4 rất dễ bị ảnh hưởng bởi sự xuất hiện của các giá trị tột cùng đặc biệt nếu chúng ta tiến hành các kiểm định giá trị trung bình. Giá trị tột cùng này có thể là hậu quả của lỗi nhập số liệu như 123 được nhập thay vì 13, hoặc chúng cũng có thể là các giá trị thực sự. Liệt kê các đối tượng có giá trị tột cùng bạn có thể quay trở lại và kiểm tra phiếu điều tra để xem giá trị đó có thực hay không. Nếu đó là giá trị thực bạn sẽ cần phải đưa nó vào trong phân tích của mình hoặc giải thích cẩn thận tại sao bạn lại không đưa nó vào trong phân tích (mặc dù điều này có ảnh hưởng đến tính khái quát hoá trong kết quả của bạn). Nếu không, nó có thể được chỉnh sửa trước quá trình phân tích.

Để kiểm tra các giá trị bất thường của các biến: số ngày nằm viện, điểm QoL trước và sau khi chấn thương, chúng ta có thể dùng menu lệnh Frequencies như trình bày ở phần trên. Câu lệnh này được sử dụng cho hầu hết các biến liên tục ngoài trừ dữ liệu kiểu ngày tháng (kiểu này chúng tôi sẽ giới thiệu với các bạn riêng).

- Chọn các biến như hình dưới đây:



- Vào nút Statistics... để chọn các giá trị mô tả như sau:



1. Nhấn Continue để quay trở lại hộp thoại đầu.

2. Nhấn nút Chart và chọn Histogram rồi nhấn Continue. Sau đó nhấn OK để thực hiện. Nếu bạn không nhấn OK mà nhấn Paste thì SPSS sẽ chép toàn bộ những gì bạn vừa thực hiện bằng menu lệnh thành syntax.

SPSS cho phép bạn kiểm tra tất cả các giá trị bất thường của các biến liên tục cùng một lần sử dụng cú pháp sau. Mỗi một tên biến cần được đưa ra một cách chính xác như nó xuất hiện trong bộ số liệu.

SYNTAX

FREQUENCIES VARIABLES = ageround q9 qol_bef qol_aft/STA= MEAN STDDEV MEDIAN MIN MAX SKEW KURT/HISTOGRAM.

Cú pháp này yêu cầu các tần số của các biến liên tục được đưa ra trong bảng trên, và cũng yêu cầu vẽ biểu đồ cột liên tục cho các biến này. Bạn có thể thay tên biến (ở cú pháp trên tên biến là những chữ nhỏ) bằng những tên biến khác từ bộ số liệu của bạn.

KẾT QUẢ

Frequencies

Statistics

		age (rounded)	hospital day	General quality of life before injury	general quality of life after injury
N	Valid	1721	810	1692	1693
	Missing	0	911	29	28
Mean		29.83	11.15	60.4054	54.6669
Median		28.00	7.00	60.0000	55.0000
Std. Deviation		15.245	15.689	7.67448	9.99864
Skewness		.266	4.611	.125	-.046
Kurtosis		-.688	35.761	.123	.059
Minimum		0	1	37.00	20.00
Maximum		65	200	88.00	88.00

Frequency Table

age (rounded)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	2	.1	.1	.1
	1	1	.1	.1	.2
	2	9	.5	.5	.7
	3	19	1.1	1.1	1.8
	4	24	1.4	1.4	3.2
	5	14	.8	.8	4.0
	6	29	1.7	1.7	5.7
	7	21	1.2	1.2	6.9
	8	38	2.2	2.2	9.1
	9	10	.6	.6	9.7
	10	19	1.1	1.1	10.8
	11	19	1.1	1.1	11.9
	12	22	1.3	1.3	13.2
	13	23	1.3	1.3	14.5
	14	27	1.6	1.6	16.1
	15	25	1.5	1.5	17.5
	16	49	2.8	2.8	20.4
	17	37	2.1	2.1	22.5
	18	38	2.2	2.2	24.8
19	48	2.8	2.8	27.5	



hospital_day

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	140	8.1	17.3	17.3
	2	67	3.9	8.3	25.6
	3	75	4.4	9.3	34.8
	4	37	2.1	4.6	39.4
	5	40	2.3	4.9	44.3
	6	6	.3	.7	45.1
	7	144	8.4	17.8	62.8
	8	8	.5	1.0	63.8
	9	3	.2	.4	64.2
	10	51	3.0	6.3	70.5
	11	4	.2	.5	71.0
	12	6	.3	.7	71.7
	13	1	.1	.1	71.9
	14	11	.6	1.4	73.2
	15	64	3.7	7.9	81.1
	16	6	.3	.7	81.9
	17	3	.2	.4	82.2

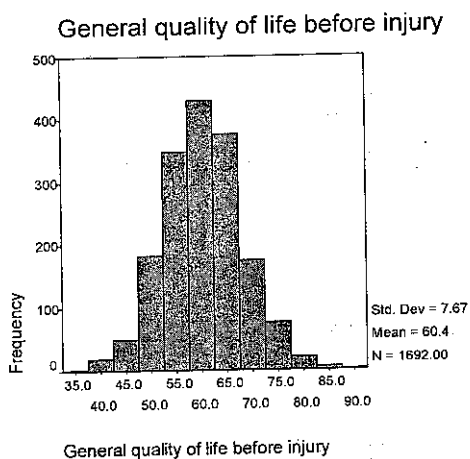
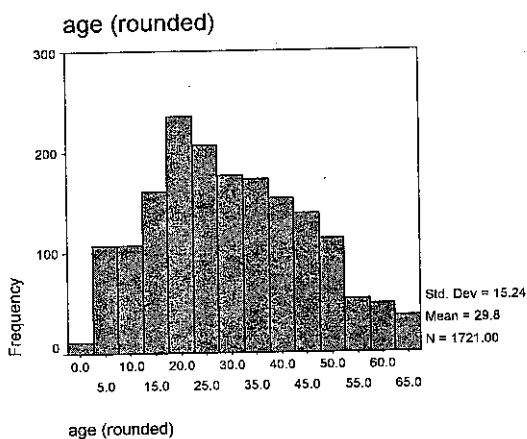
General quality of life before injury

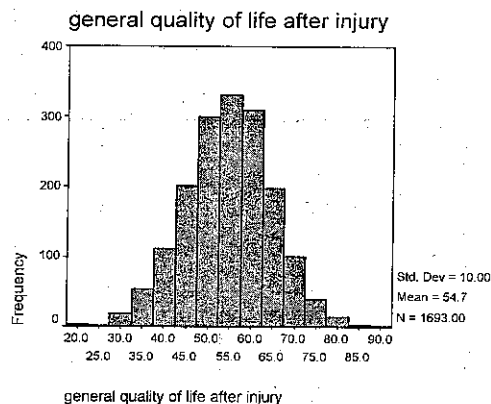
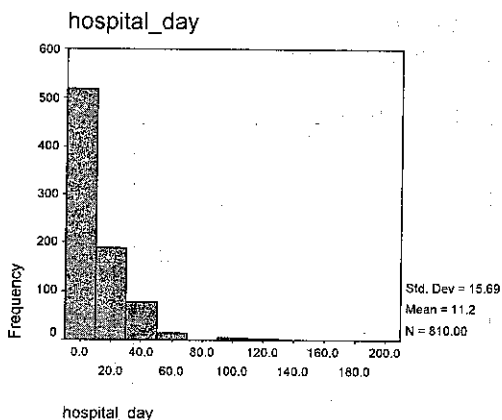
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	37.00	1	.1	.1	.1
	38.00	1	.1	.1	.1
	39.00	3	.2	.2	.3
	40.00	2	.1	.1	.4
	41.00	8	.5	.5	.9
	42.00	5	.3	.3	1.2
	43.00	2	.1	.1	1.3
	44.00	7	.4	.4	1.7
	45.00	5	.3	.3	2.0
	46.00	13	.8	.8	2.8
	47.00	25	1.5	1.5	4.3
	48.00	17	1.0	1.0	5.3
	49.00	31	1.8	1.8	7.1
	50.00	38	2.2	2.2	9.3
	51.00	53	3.1	3.1	12.5
	52.00	45	2.6	2.7	15.1
	53.00	50	2.9	3.0	18.1

general quality of life after injury

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	20.00	2	.1	.1	.1
	22.00	1	.1	.1	.2
	26.00	1	.1	.1	.2
	27.00	1	.1	.1	.3
	28.00	1	.1	.1	.4
	29.00	3	.2	.2	.5
	30.00	5	.3	.3	.8
	31.00	5	.3	.3	1.1
	32.00	5	.3	.3	1.4
	33.00	8	.5	.5	1.9
	34.00	10	.6	.6	2.5
	35.00	8	.5	.5	3.0
	36.00	13	.8	.8	3.7
	37.00	14	.8	.8	4.5
	38.00	18	1.0	1.1	5.6
	39.00	20	1.2	1.2	6.8
	40.00	17	1.0	1.0	7.8
	41.00	25	1.5	1.5	9.3
	42.00	32	1.9	1.9	11.2

Histogram





Kiểm tra ngày tháng để phát hiện sai ngày tháng

Bạn đang tìm những ngày tháng không hiệu lực, ví dụ ngày phỏng vấn nằm ngoài thời gian thực hiện nghiên cứu. Bạn nên xác định bất kỳ một giá trị ngày tháng không phù hợp nào và đối chiếu với phiếu gốc để xác minh lại. Trong trường hợp dưới đây bạn có thể muốn kiểm tra ngày 08/12/2001 là ngày phỏng vấn khi thời gian phỏng vấn từ tháng 8 đến tháng 11. Để tìm ra những đối tượng nghiên cứu được phỏng vấn ngày 08/12/2001 bạn không cần thiết phải tìm trong bộ số liệu của mình. Thay vì làm như vậy, bạn yêu cầu SPSS 'liệt kê' những trường hợp có ngày phỏng vấn là 8/12/2001, khi có danh sách những mã phiếu này bạn sẽ dễ dàng nhận ra chúng. Phần thứ hai của cú pháp yêu cầu SPSS thực hiện việc này.

Syntax

```
FREQUENCIES VARIABLES = date.
```

```
TEMPORARY.
```

```
SELECT IF (date = date.mdy(12,08,2001)).
```

```
LIST h_id date.
```

Output

Frequencies

Statistics

date of interview

N	Valid	1796
	Missing	10

date of interview.

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 11-AUG-2001	8	.3	.4	.4
13-AUG-2001	12	.7	.7	1.1
14-AUG-2001	50	2.9	2.9	4.0
15-AUG-2001	48	2.8	2.8	6.8
16-AUG-2001	40	2.3	2.3	9.1
17-AUG-2001	27	1.6	1.6	10.7

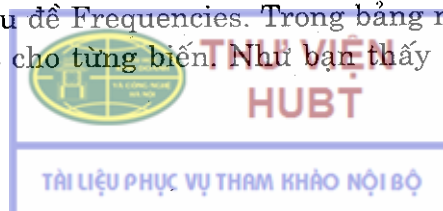
through to

29-SEP-2001	1	.1	.1	94.2
30-SEP-2001	9	.5	.5	94.7
01-OCT-2001	4	.2	.2	95.0
02-OCT-2001	1	.1	.1	95.0
03-OCT-2001	32	1.9	1.9	96.9
04-OCT-2001	8	.5	.5	97.4
05-OCT-2001	16	.9	.9	98.3
06-OCT-2001	3	.2	.2	98.5
07-OCT-2001	22	1.3	1.3	99.8
08-OCT-2001	3	.2	.2	99.9
08-DEC-2001	1	.1	.1	100.0
Total	1711	99.4	100.0	
Missing System	10	.6		
Total	1721	100.0		

Kiểm tra số lượng thông tin bị mất

Điều quan trọng là kiểm tra số lượng những trường hợp mất thông tin trong từng biến. Các đối tượng phỏng vấn chỉ có trong phân tích nếu họ đưa ra những câu trả lời có giá trị mà bạn đang xem xét. Ví dụ bạn có thể có 1721 đối tượng trong bộ số liệu của bạn nhưng thực tế chỉ có 1504 người đã trả lời cho câu hỏi về chất lượng cuộc sống, vì thế tất cả các phân tích có liên quan đến chất lượng cuộc sống chỉ được thực hiện trên 1504 người chứ không phải 1721 người. Nếu bạn có quá nhiều trường hợp bị mất số liệu thì bạn nên nghi ngờ rằng bạn đã mắc lỗi thu thập số liệu trong nghiên cứu, ví dụ những người nhận thấy chất lượng cuộc sống của họ rất kém thì gần như họ sẽ không trả lời cho câu hỏi về chất lượng cuộc sống. Bạn cần kiểm tra và tìm ra những mối liên quan giữa tính hợp lý và tính khái quát trong nghiên cứu của mình. Không có sự nhất quán trong việc xác định số liệu mất bao nhiêu là “quá nhiều”. Theo một quy ước thô, nếu số liệu bị mất với bất kỳ biến nào dưới 10% là chấp nhận được. Nếu bạn bị mất trên 10% bạn nên tham khảo một chuyên gia thống kê xem làm thế nào để đối phó với vấn đề này.

Nếu bạn nhìn vào phần kết quả ở trên, có một bảng xuất hiện ngay từ đầu của bản kết quả với tiêu đề Frequencies. Trong bảng này SPSS cho bạn thấy số lượng thông tin bị mất cho từng biến. Như bạn thấy các biến này có dưới 10%



trường hợp mất thông tin, ngoại trừ số ngày nằm viện có 911/1721 trường hợp mất thông tin. Nên nhớ giải thích kết quả của bạn thật cẩn thận. Chỉ có những người phải nằm viện chúng ta mới hỏi số ngày nằm viện, vì thế các trường hợp mất thông tin này bao gồm cả những người thật sự đã không trả lời và những người không được hỏi câu hỏi này.

Kiểm tra tính không nhất quán

Mặc dù việc kiểm tra những câu trả lời không nhất quán là một phần cần thiết trong việc làm sạch số liệu nhưng không có nguyên tắc nào cho việc xác định những mối liên quan của các câu hỏi bạn sẽ kiểm tra. Trong phạm vi nghiên cứu của mình bạn sẽ có một dự kiến về những mối liên quan có thể có giữa các câu trả lời và những mối liên quan nào là không thể. Với bộ số liệu của cuộc điều tra chấn thương giao thông quốc gia, có một số sự không nhất quán mà chúng ta có thể kiểm tra.

- i. Có ai không bị chấn thương mà lại vào viện không? (Điều này có thể đúng nhưng vẫn là hữu ích khi đối chiếu lại với phiếu gốc).
- ii. Có sự kết hợp không có thực giữa trình độ học vấn và nghề nghiệp không?
- iii. Những người đi bộ có bị phân loại vào nhóm những người khách trên xe hay là lái xe không?
- iv. Có những đứa trẻ dưới 5 tuổi mà đã đi học không?
- v. Có những đứa trẻ từ 6–9 tuổi mà lại học cao hơn tiểu học không?

Để kiểm tra sự không nhất quán bạn cần yêu cầu SPSS tìm ra bất kỳ ai có sự kết hợp của những câu trả lời bạn đã định, ví dụ tuổi nhỏ hơn 6, trình độ học vấn ở nhóm 3 (THCS) và liệt kê cho bạn.

Bạn không chỉ tìm ra những người trả lời bằng với một giá trị nào đó, mà bạn cũng có thể dùng SPSS để tìm ra những người trả lời ít hơn, nhiều hơn, hay không bằng... Các mã bạn dùng là:

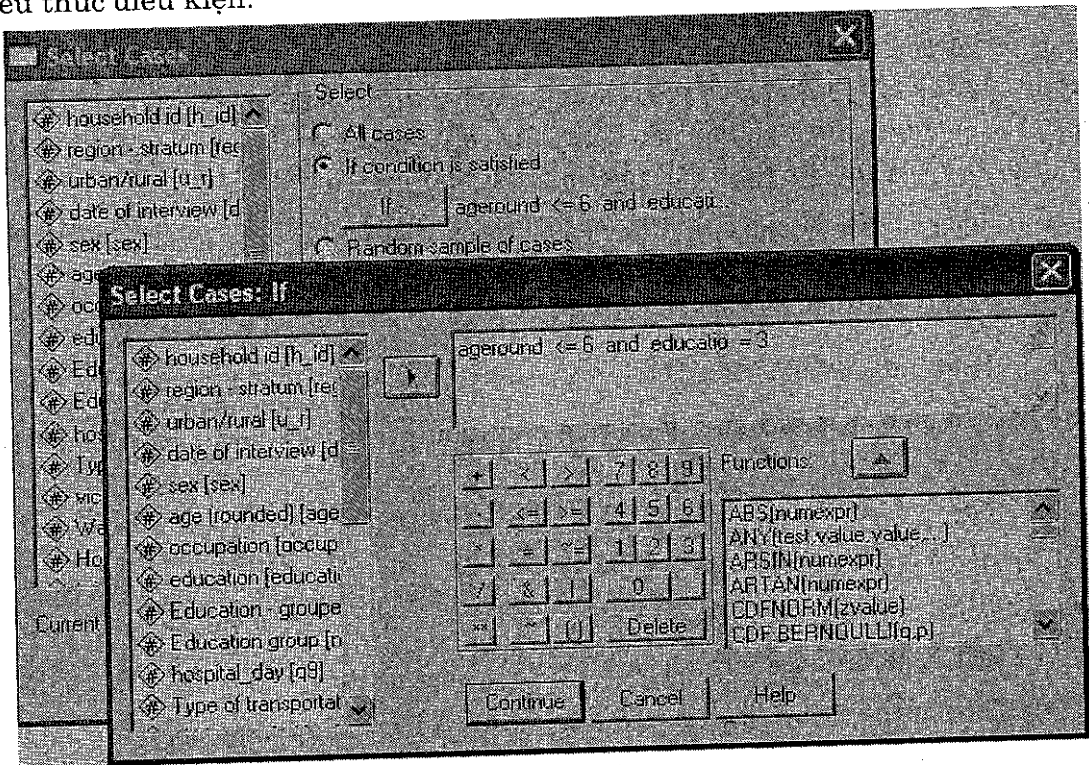
Variable EQ 0	Bằng 0
Variable NE 0	Không bằng 0
Variable GT 0	Lớn hơn 0
Variable LT 0	Nhỏ hơn 0
Variable GE 0	Lớn hơn hoặc bằng 0
Variable LE 0	Nhỏ hơn hoặc bằng 0

Tuy nhiên, chúng ta hoàn toàn có thể dùng các dấu =, <>, >, <, >=, <= thay cho các ký tự chữ (EQ, NE,...) nêu trên.

Bạn có thể thay 0 bằng bất kỳ số nào thích hợp hoặc dùng một biến để tạo nên các lời phát biểu như ngày sinh phải trước ngày tử vong.

Để tìm ra những người có những điều kiện nào đó, chẳng hạn: tìm những người có tuổi ≤ 6 và trình độ học vấn là 3 (Trung học phổ thông), bạn làm như sau:

1. Vào Data/ Select Cases..., chọn If, một cửa sổ mới hiện ra để bạn viết các biểu thức điều kiện.



2. Nhấn Continue rồi nhấn OK để thực hiện ngay hoặc nhấn Paste.

3. Chạy lệnh Frequencies cho biến h_id để liệt kê những người thoả mãn điều kiện trên.

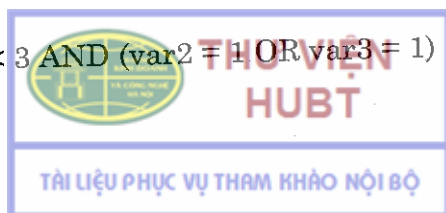
Bạn cũng có thể dùng lệnh SELECT IF, viết trực tiếp trong syntax. Dưới đây là các cấu trúc cú pháp viết trong syntax để thực hiện kiểm tra tính nhất quán của số liệu:

SYNTAX

SELECT IF dob LT dod.

Bạn có thể kết hợp các câu lệnh sử dụng từ AND và OR để tạo ra cú pháp ví dụ:

SELECT IF var1 < 3 AND (var2 = 1 OR var3 = 1)



Lệnh SELECT IF giới hạn bất kỳ lệnh nào theo sau chỉ thực hiện trên một tập hợp con của bộ số liệu bạn có. Bạn có thể làm như vậy cho bất kỳ một lệnh nào, ví dụ:

```
SELECT IF gender = 2
```

```
FREQUENCIES VARIABLES = var1
```

SPSS sẽ cho bạn tần số của biến 1 (var1) với những đối tượng mà giới tính được mã là 2.

Bạn có thể viết những câu nhắc bạn từng phần nằm trong cú pháp mà không phải là câu lệnh cho SPSS thực hiện theo. Để SPSS biết đó không phải là lệnh để nó thực hiện bạn cần bắt đầu câu với “*” và kết thúc với dấu “.” SPSS sẽ nhận ra rằng bất kỳ câu nào như vậy đều không phải là lệnh.

Dưới đây là tất cả các cú pháp cần để kiểm tra 5 sự không nhất quán có thể xác định trong bộ liệu.

Syntax

**** KIEM TRA TINH NHAT QUAN**

**** [1] lieu co trong hop nao khong co thuong tich tren nguoi vao nhap vien khong.**

```
TEMPORARY.
```

```
SELECT IF (q9 > 0 AND worst = -1).
```

```
FORMATS q9 worst (f3.0).
```

```
LIST h_id q9 worst.
```

**** [2] kiem tra tinh nhat quan cua trinh do VH va nghe nghiep.**

*** Kiem tra nhung trong hop khong nhat quan**

```
CROSSTABS TABLES = educatio BY occupatio/cell = count.
```

*** Liet ke cac ban ghi khong phu hop**

```
TEMPORARY.
```

```
SELECT IF (educatio = 1 AND occupati = 2).
```

```
LIST h_id educatio occupatio ageround sex u_r.
```

**** [3] Lieu co nan nhan bi TNGT khi di bo lai co so lieuphan lai xe hoac hanh khach?**

```
TEMPORARY.
```

```
SELECT IF (trantype = 4 AND (q41_e = 1 OR q41_e = 2)).
```

```
LIST h_id trantype q41_a q41_e.
```

** [4] Lieu co tre duoi 5 tuoi nao da di hoc khong?

TEMPORARY.

SELECT IF (ageround <= 5 AND (educatio <> 1 AND educatio <> 8)).

LIST h_id ageround educatio.

** [5] Lieu co tre tu 6-9 tuoi hoc tren cap 1 khong?

TEMPORARY.

SELECT IF (ageround >= 6 AND ageround <= 9 AND educatio <> 2).

LIST h_id ageround educatio.

KẾT QUẢ

Kết quả được đưa ra ở dưới là các kiểm tra [4] và [5]. Nếu không có đối tượng nào trả lời không nhất quán phần kết quả sẽ cho bạn thấy không có trường hợp nào được tìm ra và được liệt kê trong ví dụ thứ nhất. Nếu SPSS tìm thấy các đối tượng có câu trả lời không nhất quán nó sẽ liệt kê ra các mã phiếu cũng như thông tin khác mà bạn yêu cầu và cho bạn biết có bao nhiêu trường hợp như vậy, trong trường hợp này những đứa trẻ tuổi từ 6-9 mà học trên tiểu học được tìm ra là 9.

List

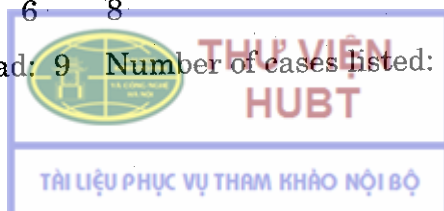
Number of cases read: 0 Number of cases listed: 0

List

H_ID AGEROUND EDUCATIO

41012008	7	8
41123022	8	3
41132055	6	8
61732173	8	8
61831086	7	8
61832013	6	8
72113047	7	8
72132065	6	8
72132065	6	8

Number of cases read: 9 Number of cases listed: 9



2.3.2. Sử dụng SPSS để quản lý số liệu.

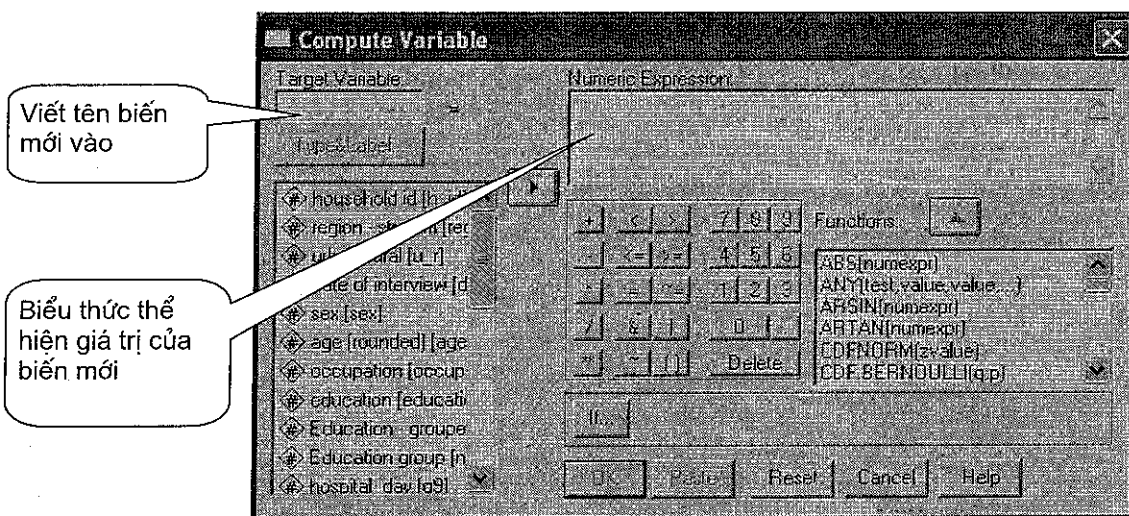
Các kiểm tra làm sạch số liệu ở trên đã được thực hiện trên các biến có câu trả lời trực tiếp trên phiếu phỏng vấn. Ví dụ, phương tiện giao thông được phân làm 5 loại là dựa vào những trả lời cho câu hỏi 7. Điều gì sẽ xảy ra nếu chúng ta quan tâm đến các đối tượng là người đi bộ hay không phải người đi bộ hơn là quan tâm đến loại xe mà họ sử dụng. Chúng ta cần phải có một biến mà chỉ có 2 phân loại chứ không phải là 5, điều này sẽ dễ dàng cho chúng ta biết đối tượng có phải là người đi bộ hay không. Bạn có thể tạo ra biến mới bằng cách hợp nhất các phân loại của biến cũ. Thao tác này là một phần của quản lý số liệu. Phần này sẽ giúp bạn biết cách làm thế nào để quản lý số liệu:

- Tạo và tính toán các biến mới.
- Chọn một tập hợp nhỏ trong các bản ghi để sử dụng.
- Mã lại các biến.

2.3.2.1. Tạo biến mới

Đôi khi bạn muốn tính toán một biến mới dựa trên các biến hiện có trong bộ số liệu. Ví dụ, chúng ta cần biết sự khác nhau giữa điểm chất lượng cuộc sống sau chấn thương với trước chấn thương ở từng đối tượng. Chúng ta có thể tính được bằng tay nhưng tính cho 1721 đối tượng sẽ tốn rất nhiều thời gian. Thay vì tính bằng tay chúng ta có thể sử dụng SPSS tính toán sự khác nhau này và đưa các giá trị vào biến mới. Trong ví dụ này biến mới được gọi là diff.

1. Để dùng thực đơn lệnh, bạn vào Transform/ Compute...



Bạn đặt tên biến mới và tính giá trị cho biến đó, bạn có thể dùng các hàm tính toán ở danh sách bên dưới Function và kết hợp dùng các toán tử.

Syntax

* định nghĩa biến mới=hệ số của QoL trước và sau chấn thương

`compute diff = qol_aft - qol_bef.`

`freq var = qol_bef qol_aft diff/sta=mean median min max/histogram.`

Dòng bắt đầu bằng COMPUTE: yêu cầu SPSS tính một biến mới.

Dòng bắt đầu bằng FREQ: yêu cầu một số thống kê tóm tắt cho hai biến gốc và biến mới giúp bạn có thể kiểm tra xem biến mới có được tính đúng không.

2.3.2.2. Chọn một tập hợp nhỏ các bản ghi

Đôi khi bạn chỉ muốn xem một nhóm đối tượng nào đó chứ không phải là tất cả. Điều này sẽ phụ thuộc vào câu hỏi nghiên cứu của bạn là gì. Ví dụ, nếu bạn viết báo cáo cho một tổ chức về sức khỏe của trẻ em, bạn có thể chỉ muốn xem xét những đối tượng là trẻ em trong bộ số liệu của bạn. Nếu bạn viết báo cáo về an toàn xe bạn có thể chỉ cần quan tâm đến những đối tượng chấn thương khi đi xe và loại trừ những đối tượng đi bộ ra khỏi phân tích. Trong bộ số liệu biến loại phương tiện phân làm 5 loại: 1 = ô tô; 2 = xe đạp; 3 = xe máy; 4 = người đi bộ và 5 = khác. Bạn cần chọn những đối tượng mà loại phương tiện không bằng 4, có nghĩa không phải người đi bộ. SPSS có thể 'lọc' và chọn các đối tượng mà bạn yêu cầu. Để làm được điều này dùng cú pháp ở dưới. Nhớ sử dụng phần thứ hai của cú pháp ở dưới để loại bỏ sự lọc này và chọn lại tất cả các đối tượng cho những phân tích về sau.

Trong phần này, bạn có thể sử dụng Data/ Select Cases... như đã hướng dẫn trong phần 2.3.1

Syntax

* Loại bỏ các trường hợp chấn thương khi đi bộ vào phân tích

`COMPUTE filter_$=(trantype ne 4).`

`VARIABLE LABEL filter_$ 'trantype ne 4 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .`

* Chọn toàn bộ các trường hợp

`FILTER OFF.`



USE ALL.

EXECUTE .

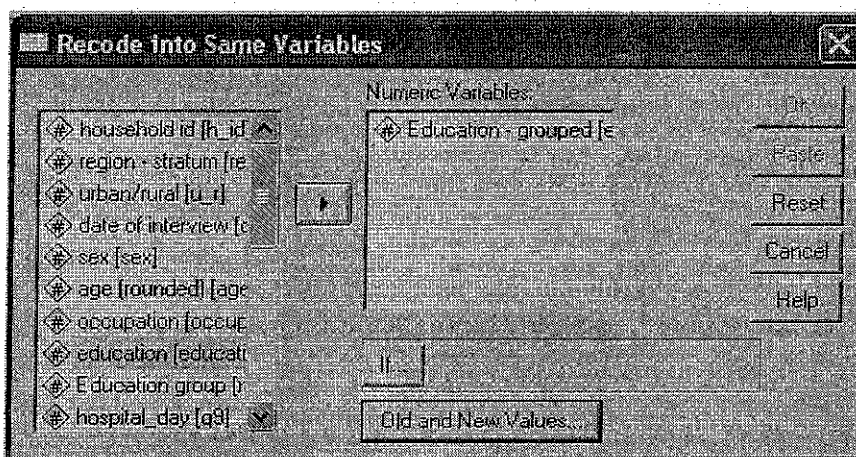
2.3.2.3. Mã hoá lại các biến

Mã hoá lại một biến phân loại

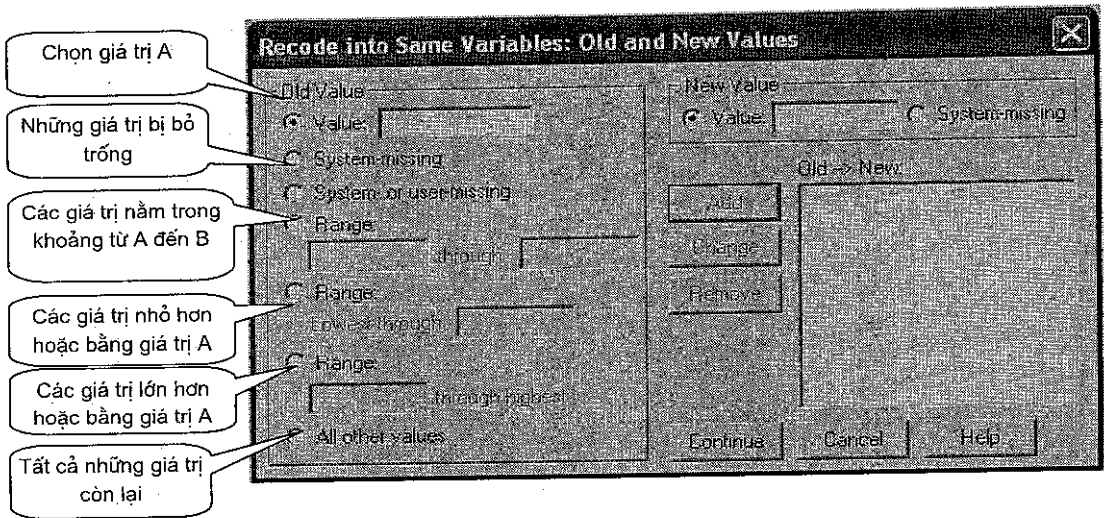
Trong phiếu điều tra chấn thương giao thông quốc gia câu 4 hỏi về trình độ học vấn, và câu 8 hỏi về vị trí của chấn thương trầm trọng nhất. Trong phân tích của mình các nhà nghiên cứu không muốn có quá nhiều phân loại trình độ học vấn và họ chỉ quan tâm các chấn thương ở đầu/cột sống trong so sánh với các vị trí chấn thương khác. Họ muốn tạo ra các biến mới mà có ít phân loại hơn. Để làm được điều này họ có thể tạo nên các biến mới bằng cách gộp các phân loại hiện tại, và kết hợp một số phân loại với nhau, tạo ra một biến trình độ học vấn mới edgrp với 4 loại, và một biến chấn thương ở đầu/cột sống mới chỉ với 2 loại.

Dùng menu lệnh: **Transform/ Recode:**

Nếu chọn Into same variables...: mã hoá lại giá trị của một biến và kết quả mới sẽ được lưu đè lên chính biến đó. Khi đó, cửa sổ sau hiện ra, bạn chọn biến cần nhóm, chẳng hạn chọn biến Education:



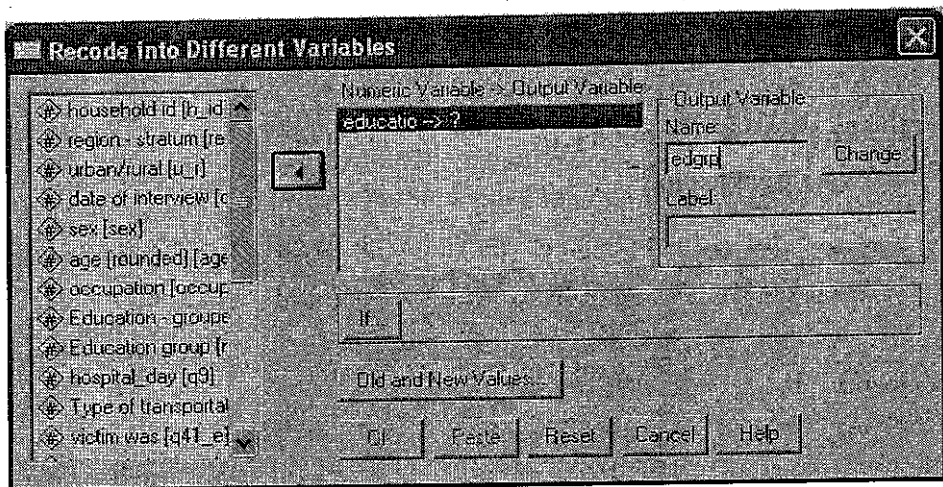
Nhấn nút Old and New Values... để chọn giá trị nhóm, cửa sổ sau hiện ra:



1. Bên phần Old Value, chọn những giá trị sẽ được phân vào một nhóm, sau đó nhập giá trị mới, đại diện cho cả nhóm đó vào ô Value ở phần New value.

2. Nhấn nút Add và tiếp tục làm với các phân nhóm tiếp theo.

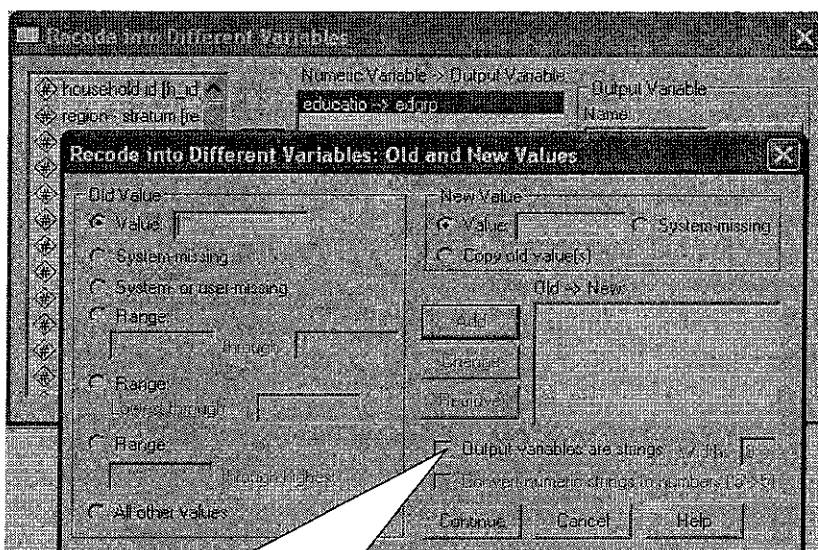
Nếu chọn Into different variables nghĩa là bạn sẽ thực hiện mã hoá một biến nhưng kết quả mã hoá được lưu vào một biến khác và giá trị “gốc” vẫn được giữ nguyên.



1. Chọn biến cần mã hoá từ danh sách các biến và kích mũi tên, ở đây chẳng hạn chọn biến education.

2. Nhập tên biến mới vào hộp Name và nhấn nút Change.

3. Nhấn nút Old and New Values... để thực hiện mã hoá (phân nhóm) cho các giá trị. Quá trình này cũng được thực hiện giống như phần trên.



Nếu bạn muốn mã hoá thành những mã kí tự thì chọn vào đây. Biến mới sẽ là biến kiểu string thay vì kiểu Number (số)

4. Hộp Old → New sẽ lưu danh sách các phân nhóm mà bạn vừa làm, từ danh sách này, bạn có thể kiểm tra xem mình đã phân nhóm đúng chưa. Bạn cũng có thể sửa lại bằng cách chọn nhóm trong danh sách đó và kích vào nút Change, kích vào nút Remove nếu bạn muốn xoá bỏ.

5. Nhấn Continue để khẳng định việc mã hoá bạn vừa làm rồi nhấn OK để thực hiện.

SYNTAX

**** Định nghĩa biến mới edgrp**

COMPUTE edgrp=educatio.

RECODE edgrp (8=1) (1,2=2) (3,4=3) (5 thru 7=4) (else=-1).

VAR LABELS edgrp 'Education - grouped'.

VALUE LABELS edgrp 1 'Children' 2 'Less than secondary' 3 'Secondary' 4 'More than secondary'.

MISSING VALUES edgrp (-1).

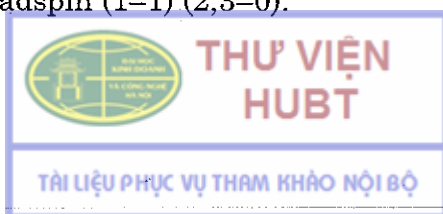
EXECUTE.

FREQUENCIES VARIABLES = educatio edgrp.

**** Định nghĩa biến mới headspin (đau và xương song)**

COMPUTE headspin=worst.

RECODE headspin (1=1) (2,3=0).



VAR LABELS headspin 'Injury to head or spine'.

VALUE LABELS headspin 0 'Not injured at these sites' 1 'Injured'.

MISSING VALUES headspin (-1).

EXECUTE.

FREQUENCIES VARIABLES = worst headspin.

Dòng COMPUTE yêu cầu SPSS tạo ra biến mới mà bạn có thể nhóm lại (không bao giờ nhóm biến gốc vì bạn có thể mất các số liệu hiện có).

Dòng RECODE cho SPSS biết nhóm biến mới này như thế nào ví dụ như phân loại của biến cũ sẽ chuyển sang các phân loại của biến mới như thế nào.

Dòng VAR LABELS: gán nhãn cho biến mới để bạn có thể biết nó là biến gì.

Dòng VALUE LABELS: gán các nhãn cho mã của biến mới để bạn có thể biết các mã đó có nghĩa là gì.

Dòng MISSING VALUES: cho SPSS biết mã các giá trị bị mất là gì.

Dòng cuối cùng FREQUENCIES: nên được chạy để kiểm tra rằng mã của biến mới đã thực hiện đúng. Đây không phải là một kiểm định nhưng đôi khi bạn nên xem xét kỹ kiểm tra các số phân loại của biến mới là đúng và phân bố các trường hợp là ổn.

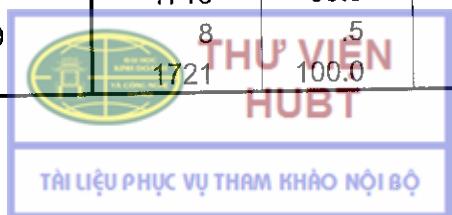
Ví dụ, bản ghi như sau là đúng:

Most severe injury

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid -1.00	62	3.6	3.6	3.6
head/spine	624	36.3	36.3	39.9
torso	379	22.0	22.0	61.9
limbs	656	38.1	38.1	100.0
Total	1721	100.0	100.0	

Injury to head or spine

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid Not injured at these sites	1089	63.3	63.6	63.6
Injured	624	36.3	36.4	100.0
Total	1713	99.5	100.0	
Missing -1.00	8	.5		
Total	1721	100.0		



Phân loại một biến liên tục

Để phân loại một biến liên tục thay vì định rõ các giá trị mới cho các phân loại hiện tại bạn sẽ cần cho SPSS biết sự sắp xếp của biến liên tục tương ứng với các phân loại bạn muốn trong biến mới. Ví dụ, của chất lượng cuộc sống (mã là 0) được xác định là 50 hoặc nhỏ hơn, với 51 điểm hoặc cao hơn được xác định là điểm tương xứng của chất lượng cuộc sống (mã là 1). Nếu hai biến liên tục có thể được gộp lại theo một cách giống hệt nhau thì bạn có thể thực hiện trên cả hai biến trong cùng một lần.

Để mã hoá các biến liên tục, bạn có thể dùng menu lệnh như trình bày ở trên, hoặc sử dụng syntax như sau.

Syntax

*Định nghĩa 2 biến mới QoL trước và QoL sau.

```
freq var = qol_bef qol_aft/format=notable/sta=min max.
```

```
COMPUTE qolbefg = qol_bef.
```

```
COMPUTE qolaftg = qol_aft.
```

```
RECODE qolbefg qolaftg (0 thru 50=0) (51 thru 100=1).
```

```
VALUE LABELS qolbefg qolaftg 0 'Suboptimal QOL' 1 'Adequate QOL'.
```

Dòng COMPUTE: cho SPSS biết tạo ra hai biến mới mà bạn có thể nhóm được (không bao giờ nhóm các biến gốc vì bạn có thể bị mất các số liệu hiện có).

Dòng RECODE: cho SPSS biết nhóm biến mới này như thế nào.

Dòng VALUE LABELS: gắn các nhãn cho mã của biến mới cho phép bạn biết từng mã có nghĩa là gì.

Tóm tắt

Điều quan trọng là phải dành thời gian thích đáng cho việc nhập số liệu, mã số liệu, làm sạch số liệu và quản lý bộ số liệu của bạn. Nếu bạn làm những việc này tốt bạn sẽ tiết kiệm được thời gian và các vấn đề nảy sinh về sau khi bạn bắt đầu phân tích số liệu. Giữ một bản ghi chính xác tất cả những gì bạn đã làm trong phần này để bạn có thể quay lại kiểm tra bất kỳ vấn đề gì xuất hiện sau này. Một cách tốt để giữ các bản ghi này là ghi lại tất cả các cú pháp mà bạn đã viết. Đó cũng là một sáng kiến tốt vì bạn có thể chỉnh sửa nó và tiếp tục dùng cho những nghiên cứu tiếp theo. Một khi số liệu của bạn được nhập và làm sạch bạn đã sẵn sàng cho bước tiếp theo, đó là phân tích số liệu. Hai bài 3 và 4 tiếp theo sẽ trình bày về quá trình phân tích số liệu.

Bài số

3

PHÂN TÍCH THỐNG KÊ MÔ TẢ

Mục tiêu

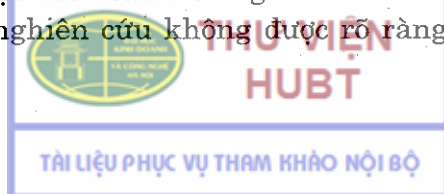
1. Lựa chọn được các biểu đồ/diagram thích hợp cho việc mô tả các loại biến số và mối liên quan.
2. Đưa ra các lý do của sự lựa chọn đó.
3. Hiểu được các giá trị liên quan đến từng tóm tắt.
4. Sử dụng được phần mềm SPSS để phân tích thống kê và vẽ đồ thị.

3.1. Giới thiệu

Phân tích số liệu bao gồm việc tóm tắt và so sánh các số liệu định lượng để trả lời các câu hỏi nghiên cứu – để mô tả các đo lường thay đổi như thế nào và xác minh mối có liên quan gì giữa các biến. Trong cuốn Thống kê Y tế Công cộng phần Thống kê cơ bản, bạn đã học về các tiếp cận cơ bản để tóm tắt thống kê và kiểm định giả thuyết. Chúng ta có khá nhiều kiểm định thống kê và một số kiểm định trong số đó có những đặc điểm khá là giống nhau và đôi khi sự tương đồng này dẫn đến những sự nhầm lẫn của chúng ta.

Chúng ta sẽ sử dụng những kiểm định nào và vào lúc nào? Quá trình lựa chọn kiểm định thống kê thích hợp cho một bộ số liệu chính là kế hoạch phân tích của bạn. Việc có một kế hoạch phân tích chi tiết, rõ ràng sẽ giúp bạn tiết kiệm rất nhiều thời gian và tránh những sai sót về sau này.

Bài này giới thiệu cho bạn những khái niệm của kế hoạch phân tích. Thời điểm lý tưởng để đưa ra một kế hoạch phân tích là trong giai đoạn thiết kế nghiên cứu, khi mà nhóm nghiên cứu đang lập kế hoạch đo lường cái gì, ai và khi nào. Các câu hỏi nghiên cứu cần thiết phải được trả lời trong nghiên cứu và chúng ta sẽ không thể lập kế hoạch phân tích số liệu nếu chúng ta thiếu câu hỏi nghiên cứu được đưa ra rõ ràng. Nếu nghiên cứu này là do bạn thiết kế và tự thu thập số liệu thì việc biết câu hỏi nghiên cứu là một việc đơn giản. Tuy nhiên, đôi khi câu hỏi nghiên cứu không được rõ ràng lắm khi bạn chỉ là một



thành viên của nhóm nghiên cứu và bạn không phải là người lãnh đạo nhóm. Trong trường hợp này, nếu bạn là người phân tích bạn phải thảo luận câu hỏi nghiên cứu với các thành viên khác của nhóm nghiên cứu.

3.2. Tiến trình của kế hoạch phân tích

Phân tích thống kê một bộ số liệu không khó nếu số liệu đã được làm sạch và chuẩn bị thích đáng cho việc phân tích (xem bài 2) và các giả thuyết nghiên cứu được xác định một cách rõ ràng (xem bài 1). Phần khó khăn của phân tích số liệu là xác định câu hỏi nghiên cứu một cách rõ ràng, phần còn lại là việc chúng ta làm theo một “công thức”. Quyển sách này và bài này sẽ cung cấp cho bạn một “công thức” cho hầu hết các phân tích thống kê cơ bản thông thường mà bạn sẽ thực hiện trong các nghiên cứu sức khỏe.

Bạn nên chuẩn bị một kế hoạch về những việc bạn sẽ làm để tóm tắt và phân tích bộ số liệu. Có rất nhiều câu hỏi bạn cần trả lời để chuẩn bị kế hoạch của mình.

1. Câu hỏi nghiên cứu chỉ liên quan đến mô tả số liệu hay nó yêu cầu kiểm định giả thuyết?

Nếu chỉ mô tả số liệu, tiếp tục theo câu hỏi 3 ở dưới. Nếu không:

2. Những giả thuyết khoa học nào được bao hàm trong câu hỏi nghiên cứu? Một giả thuyết kiểm định bao gồm cả giả thuyết không (H_0) và đối thuyết (H_1). Nhưng bạn sẽ thấy, thường bao giờ cũng có nhiều hơn một giả thuyết khoa học từ một câu hỏi nghiên cứu.

3. Cho từng mục đích mô tả hoặc các giả thuyết kiểm định thực hiện, hãy:

(i) LIỆT KÊ CÁC BIẾN

- Xác định biến phụ thuộc và các biến độc lập.
- Xác định loại biến (biến liên tục/khoảng chia hoặc phân loại).

(ii) TÓM TẮT CÁC BIẾN VÀ CÁC MỐI LIÊN QUAN

- Sử dụng các thông tin từ (i), và chuyển đến các bảng 1 (tr.193) và bảng 2 (tr.194), chọn xem bạn sẽ làm thế nào để tóm tắt thống kê hoặc mối liên quan giữa hai biến.
- Đưa ra một bảng ‘giả’ mô tả các kết quả cho mối liên quan này trong báo cáo cuối cùng của bạn.
- Nếu chỉ mô tả, thì kế hoạch phân tích của bạn đã hoàn thành. Nếu không, cho mỗi giả thuyết được kiểm định.

Bài số

3

PHÂN TÍCH THỐNG KÊ MÔ TẢ

Mục tiêu

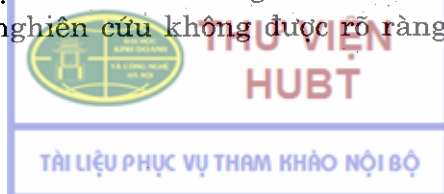
1. Lựa chọn được các biểu đồ/đồ thị thích hợp cho việc mô tả các loại biến số và mối liên quan.
2. Đưa ra các lý do của sự lựa chọn đó.
3. Hiểu được các giá định liên quan đến từng tóm tắt.
4. Sử dụng được phần mềm SPSS để phân tích thống kê và vẽ đồ thị.

3.1. Giới thiệu

Phân tích số liệu bao gồm việc tóm tắt và so sánh các số liệu định lượng để trả lời các câu hỏi nghiên cứu – để mô tả các đo lường thay đổi như thế nào và xác minh mối có liên quan gì giữa các biến. Trong cuốn Thống kê Y tế Công cộng phần Thống kê cơ bản, bạn đã học về các tiếp cận cơ bản để tóm tắt thống kê và kiểm định giả thuyết. Chúng ta có khá nhiều kiểm định thống kê và một số kiểm định trong số đó có những đặc điểm khá là giống nhau và đôi khi sự tương đồng này dẫn đến những sự nhầm lẫn của chúng ta.

Chúng ta sẽ sử dụng những kiểm định nào và vào lúc nào? Quá trình lựa chọn kiểm định thống kê thích hợp cho một bộ số liệu chính là kế hoạch phân tích của bạn. Việc có một kế hoạch phân tích chi tiết, rõ ràng sẽ giúp bạn tiết kiệm rất nhiều thời gian và tránh những sai sót về sau này.

Bài này giới thiệu cho bạn những khái niệm của kế hoạch phân tích. Thời điểm lý tưởng để đưa ra một kế hoạch phân tích là trong giai đoạn thiết kế nghiên cứu, khi mà nhóm nghiên cứu đang lập kế hoạch đo lường cái gì, ai và khi nào. Các câu hỏi nghiên cứu cần thiết phải được trả lời trong nghiên cứu và chúng ta sẽ không thể lập kế hoạch phân tích số liệu nếu chúng ta thiếu câu hỏi nghiên cứu được đưa ra rõ ràng. Nếu nghiên cứu này là do bạn thiết kế và tự thu thập số liệu thì việc biết câu hỏi nghiên cứu là một việc đơn giản. Tuy nhiên, đôi khi câu hỏi nghiên cứu không được rõ ràng lắm khi bạn chỉ là một



thành viên của nhóm nghiên cứu và bạn không phải là người lãnh đạo nhóm. Trong trường hợp này, nếu bạn là người phân tích bạn phải thảo luận câu hỏi nghiên cứu với các thành viên khác của nhóm nghiên cứu.

3.2. Tiến trình của kế hoạch phân tích

Phân tích thống kê một bộ số liệu không khó nếu số liệu đã được làm sạch và chuẩn bị thích đáng cho việc phân tích (xem bài 2) và các giả thuyết nghiên cứu được xác định một cách rõ ràng (xem bài 1). Phần khó khăn của phân tích số liệu là xác định câu hỏi nghiên cứu một cách rõ ràng, phần còn lại là việc chúng ta làm theo một “công thức”. Quyển sách này và bài này sẽ cung cấp cho bạn một “công thức” cho hầu hết các phân tích thống kê cơ bản thông thường mà bạn sẽ thực hiện trong các nghiên cứu sức khỏe.

Bạn nên chuẩn bị một kế hoạch về những việc bạn sẽ làm để tóm tắt và phân tích bộ số liệu. Có rất nhiều câu hỏi bạn cần trả lời để chuẩn bị kế hoạch của mình.

1. Câu hỏi nghiên cứu chỉ liên quan đến mô tả số liệu hay nó yêu cầu kiểm định giả thuyết?

Nếu chỉ mô tả số liệu, tiếp tục theo câu hỏi 3 ở dưới. Nếu không:

2. Những giả thuyết khoa học nào được bao hàm trong câu hỏi nghiên cứu? Một giả thuyết kiểm định bao gồm cả giả thuyết không (H_0) và đối thuyết (H_1). Nhưng bạn sẽ thấy, thường bao giờ cũng có nhiều hơn một giả thuyết khoa học từ một câu hỏi nghiên cứu.

3. Cho từng mục đích mô tả hoặc các giả thuyết kiểm định thực hiện, hãy:

(i) LIỆT KÊ CÁC BIẾN

- Xác định biến phụ thuộc và các biến độc lập.
- Xác định loại biến (biến liên tục/khoảng chia hoặc phân loại).

(ii) TÓM TẮT CÁC BIẾN VÀ CÁC MỐI LIÊN QUAN

- Sử dụng các thông tin từ (i), và chuyển đến các bảng 1 (tr.193) và bảng 2 (tr.194), chọn xem bạn sẽ làm thế nào để tóm tắt thống kê hoặc mối liên quan giữa hai biến.
- Đưa ra một bảng ‘giả’ mô tả các kết quả cho mối liên quan này trong báo cáo cuối cùng của bạn.
- Nếu chỉ mô tả, thì kế hoạch phân tích của bạn đã hoàn thành. Nếu không, cho mỗi giả thuyết được kiểm định.

(iii) CHỌN MỘT KIỂM ĐỊNH THỐNG KÊ

- Sử dụng các bảng 1 (tr.193) và bảng 2 (tr.194), với các thông tin từ (i) và (ii) ở trên, chọn hầu hết các kiểm định thống kê phù hợp.
- Kiểm tra các giả định cho kiểm định này (xem mục 4.7).
- Lựa chọn cuối cùng kiểm định dựa trên giả định có được thoả mãn hay không.

(iv) PHIÊN GIẢI CÁC KẾT QUẢ THU ĐƯỢC

- Lựa chọn mức ý nghĩa thống kê sẽ được dùng để kiểm định giả thuyết.
- Viết ra những gì bạn muốn nói về các kết quả trong báo cáo cuối cùng như thế nào nếu người đọc báo cáo của bạn là người không có chuyên môn sâu về thống kê (giả sử rằng bạn đã hoàn thành phần phân tích và tìm thấy kết quả có ý nghĩa thống kê).
- Gộp cả việc đưa ra kiểm định thống kê nào bạn đã chọn và lý do tại sao.

Phần còn lại của bài này dành cho những khái niệm của kế hoạch phân tích giúp bạn thấy một phân tích bao gồm nhiều kiểm định thống kê để trả lời cho một câu hỏi nghiên cứu. Bài này cũng mô tả cách sử dụng phần mềm thống kê SPSS để thực hiện các phân tích thống kê thông thường mà bạn cần cho việc phân tích mô tả một bộ số liệu.

3.3. Các câu hỏi nghiên cứu từ bộ số liệu mẫu

Bài 2 đã giới thiệu với bạn khái niệm về quản lý số liệu và giới thiệu một bộ số liệu từ Nghiên cứu chấn thương giao thông quốc gia. Bạn sẽ nhớ rằng chủ đề nghiên cứu là

Trong số những người bị chấn thương giao thông năm 2001, những tác động nào của chấn thương có ảnh hưởng đến chất lượng cuộc sống?

Nhóm nghiên cứu đã chuyển chủ đề nghiên cứu thành những câu hỏi nghiên cứu cụ thể:

1. Mô tả sơ lược yếu tố xã hội – dân số (giới tính, tuổi, trình độ học vấn, nghề nghiệp) của những đối tượng bị chấn thương giao thông.
2. Mô tả sơ lược về tuổi và giới của mẫu trong các vùng nghiên cứu.
3. Mô tả điểm chất lượng cuộc sống trước chấn thương, và xem nó có bị ảnh hưởng của tuổi hay không.
4. Để xác minh tính đại diện của mẫu nghiên cứu đối với quần thể chung, từ đó có thể biết kết quả nghiên cứu có khái quát được cho quần thể hay không.

H_0 : Điểm trung bình của QoL trước chấn thương là tương tự như quần thể chung, là 50 điểm.



5. Xác minh mối liên quan giữa các yếu tố xã hội – dân số với chất lượng cuộc sống trước chấn thương.

H₀: Điểm trung bình của QoL trước chấn thương là như nhau ở nam và nữ.

H₀: Điểm trung bình của QoL trước chấn thương là như nhau ở tất cả các vùng nghiên cứu.

H₀: Điểm trung bình của QoL trước chấn thương là như nhau ở tất cả các trình độ học vấn.

6. Kiểm tra sự thay đổi chất lượng cuộc sống sau chấn thương:

H₀: Điểm trung bình của QoL sau chấn thương giao thông là cao hơn hoặc không thay đổi so với trước chấn thương.

H₀: Điểm QoL thấp (điểm phân loại) là như nhau trước và sau chấn thương.

7. Xác định mối liên quan giữa số ngày nằm viện giữa những người đi bộ và những người đi xe.

H₀: Trong số những người phải nằm viện, số ngày nằm viện trung bình là tương tự nhau giữa những người đi bộ và những người đi xe.

8. Loại trừ những người đi bộ, tìm hiểu mối liên quan giữa số ngày nằm viện với loại phương tiện bị tai nạn.

H₀: Loại trừ những người đi bộ, số ngày nằm viện trung bình là tương tự nhau ở các nhóm đối tượng sử dụng các phương tiện khác nhau.

9. Tìm hiểu mối liên quan giữa chất lượng cuộc sống sau chấn thương với tuổi của người bị chấn thương hoặc số ngày nằm viện.

H₀: Điểm trung bình của QoL sau chấn thương không có mối liên quan với số ngày nằm viện.

H₀: Điểm trung bình của QoL sau chấn thương không có mối liên quan với tuổi của người bị chấn thương.

10. Nhóm nghiên cứu quan tâm đến chấn thương ở đầu/cột sống. Trong năm 1997, tỷ lệ nạn nhân chấn thương giao thông có tổn thương ở đầu/cột sống là 37%. Để xác minh xem tỷ lệ này ở năm 2001 có thay đổi hay không:

H₀: Tỷ lệ các nạn nhân bị chấn thương giao thông có tổn thương ở đầu/cột sống là 37%.

11. Có ý kiến cho rằng những người đi bộ ít được bảo vệ hơn những người đi xe, nên có thể dễ bị những chấn thương nặng đặc biệt là ở đầu/cột sống.

H₀: So với những người bị chấn thương khi đi xe, tỷ lệ chấn thương ở đầu/cột sống ở những người đi bộ là tương tự hoặc thấp hơn.

12. Nhóm nghiên cứu cũng muốn cân nhắc đến sự nhận thức về sự khác nhau điểm dưới của QoL dựa trên mức độ chấn thương, và được lượng giá bằng vị trí chấn thương có ảnh hưởng lớn nhất.

H₀: Tỷ lệ nạn nhân nhận thức về điểm dưới của QoL là tương tự nhau không kể mức độ chấn thương, và được lượng giá bằng vị trí chấn thương có ảnh hưởng lớn nhất.

3.4. Kế hoạch phân tích của bộ số liệu mẫu – thống kê mô tả

Các câu hỏi ở mục 3.3 bao gồm hai loại phân tích thống kê: phân tích mô tả cho câu hỏi 1 đến 3 và thống kê suy luận cho các câu hỏi từ 4 đến 12.

Phần còn lại của bài này sẽ nói về kế hoạch phân tích bao gồm các phân tích mô tả cho các câu hỏi nghiên cứu từ 1 đến 3. Kế hoạch phân tích cho các giả thuyết được nêu ra trong các câu hỏi nghiên cứu từ 4 đến 12 sẽ được nói đến ở bài 4.

Trong quá trình chọn phân tích mô tả của một biến hoặc một mối liên quan giữa hai biến bị ảnh hưởng bởi các đo lường của biến phụ thuộc và dạng so sánh, bạn nên tham khảo bảng 1(tr.193) và bảng 2(tr.194) ở cuối bài để đưa ra sự lựa chọn về cách làm thế nào để tóm tắt và phân tích bộ số liệu của bạn. Các bảng này tóm tắt những hướng quyết định cho hầu hết các thống kê mô tả và các kiểm định thống kê cơ bản của các biến liên tục và phân loại. Bạn hãy dành thời gian xem xét nội dung của các bảng này và cân nhắc xem chúng được sử dụng như thế nào trong các ví dụ trong bài này cũng như bài 4.

Bảng 1 (tr.193) được dùng để chọn các tóm tắt và kiểm định thống kê để phân tích **biến phụ thuộc có dạng đo lường liên tục/khoảng**.

Bảng 2 (tr.194) được dùng để chọn các tóm tắt và kiểm định thống kê để phân tích biến phụ thuộc có **dạng đo lường phân loại**.

Một trong những giả định cần phải được thoả mãn cho việc tóm tắt và phân tích các biến phụ thuộc liên tục bằng giá trị trung bình là phân bố tần số của biến phải là phân bố chuẩn. Trong khi còn rất nhiều giả định khác cũng thường cần phải cân nhắc cho các dạng kiểm định thống kê khác nhau, giả định này phải được xem xét trước những phần khác trong bảng để chọn được một tóm tắt thống kê phù hợp. Các loại giả định khác sẽ cần được cân nhắc khi chọn các kiểm định thống kê cho kiểm định giả thuyết, điều này sẽ được nói đến trong bài 4. Làm thế nào để biết phân bố có phải là phân bố chuẩn hay không được mô tả trong mục 4.7, một phần dành để giải thích tất cả các giả định bạn có thể cần phải cân nhắc.

Một kế hoạch phân tích gợi ý cho câu hỏi đầu tiên trong 3 câu hỏi mô tả được đưa ra dưới đây:



3.4.1. Kế hoạch phân tích – Thống kê mô tả

Mô tả sơ lược yếu tố xã hội – dân số (giới tính, tuổi, trình độ học vấn, nghề nghiệp) của những đối tượng bị chấn thương giao thông.

Các biến: Câu hỏi này yêu cầu tóm tắt tất cả 4 loại biến về yếu tố xã hội– dân số. Chúng là các loại khác nhau; giới tính là biến nhị phân, tuổi là biến liên tục; trình độ học vấn là biến thứ hạng và nghề nghiệp là biến danh mục.

Tóm tắt: Theo bảng 1(tr.193) nếu một biến là biến phân loại (bao gồm biến danh mục, thứ hạng, nhị phân) thì số lượng và tỷ lệ nên được dùng để đưa ra một tóm tắt biến này dưới dạng số, và biểu đồ cột sẽ thích hợp để biểu diễn biến này. Nếu là biến liên tục thì lượng giá bằng trung bình và độ phân tán là thích hợp; giá trị trung bình và độ lệch chuẩn nếu phân bố của biến là phân bố chuẩn, nếu không trung vị và khoảng (giá trị cực tiểu, giá trị cực đại) là phù hợp. Biểu đồ Box-and-Whisker sẽ phù hợp với bất kỳ một biến liên tục nào, kể cả có phân bố chuẩn hay không.

Vì thế kế hoạch xuất phát từ:

1. Tính các tần số, các tỷ lệ và các biểu đồ cột cho biến giới tính, trình độ học vấn, nghề nghiệp.
2. Kiểm tra xem tuổi có phân bố chuẩn hay không.
3. Nếu tuổi là phân bố chuẩn, thì sử dụng giá trị trung bình, độ lệch chuẩn, biểu đồ Box-and-Whisker.
4. Nếu tuổi không là phân bố chuẩn, sử dụng giá trị trung vị, cực tiểu, cực đại và biểu đồ Box-and Whisker.

Mô tả sơ lược về giới tính và tuổi của mẫu trong vùng.

Câu hỏi này yêu cầu mô tả mối liên quan giữa hai biến:

- Mô tả giới trong các vùng nghiên cứu.
- Mô tả tuổi trong các vùng nghiên cứu.

Các biến: giới tính và vùng miền đều là biến phân loại vì thế yêu cầu các thống kê mô tả cho biến phân loại bằng các mối liên quan phân loại.

Tuổi là biến liên tục và vùng miền là biến phân loại, biến này yêu cầu các thống kê mô tả.

Tóm tắt: Theo bảng 2(tr.194), nếu mối liên quan giữa hai biến phân loại thì trình bày kết quả bằng bảng có tần số và tỷ lệ. Với giới tính theo vùng, chúng ta có thể có bảng như sau:



Bảng... Phân bố giới tính theo vùng

Vùng	Tần số	Tỷ lệ nam giới (%)
Tây Bắc		
Đông Bắc		
Đồng bằng sông Hồng		
Bắc Trung Bộ		
Duyên hải ven Trung Bộ		
Tây Nguyên		
Đông Nam Bộ		
Đồng bằng sông Mekong		

Chúng ta luôn để số lượng các đối tượng trong từng loại danh mục là tốt nhất.

Kế hoạch phân tích là tính một bảng chéo về số người trong vùng phân theo giới. Vì chúng ta muốn có tỷ lệ nam theo từng vùng nên yêu cầu tính các tỷ lệ theo hàng ngang.

Bảng 1(tr.193) gợi ý rằng tóm tắt các giá trị trung bình và độ lệch chuẩn trong từng vùng là một tóm tắt phù hợp bằng số cho mối liên quan giữa tuổi trong các vùng. Bảng có dạng sau:

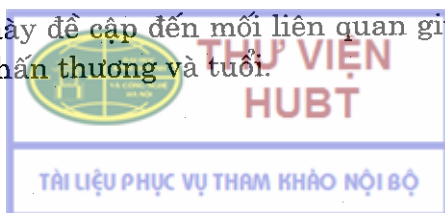
Bảng ... Phân bố tuổi theo vùng

Vùng	Số quan sát	Trung bình	Độ lệch chuẩn
Tây Bắc			
Đông Bắc			
Đồng bằng sông Hồng			
Bắc Trung Bộ			
Duyên hải ven Trung Bộ			
Tây Nguyên			
Đông Nam Bộ			
Đồng bằng sông Mekong			

Kế hoạch này có nghĩa là yêu cầu tính các giá trị trung bình và các độ lệch chuẩn trong vùng và sử dụng biểu đồ box-and-whisker để phiên giải kết quả.

Mô tả về điểm chất lượng cuộc sống trước chấn thương và xem xét điểm này có bị ảnh hưởng bởi tuổi hay không.

Các biến: Câu hỏi này đề cập đến mối liên quan giữa hai biến liên tục, chất lượng cuộc sống trước chấn thương và tuổi.



Mô tả: Từ bảng 1(tr.193) ta thấy biểu đồ phù hợp nhất dùng để mô tả mối quan hệ này là biểu đồ chấm điểm (scatter) và nếu quan hệ này có dạng tuyến tính thì hệ số tương quan là mô tả dạng số phù hợp nhất. Nếu hai biến là phân bố chuẩn thì dùng hệ số tương quan Pearson, nếu không nên sử dụng hệ số tương quan Spearman.

Thực hiện một kế hoạch phân tích

Mô tả thống kê bộ số liệu có thể dùng dưới dạng số hoặc biểu đồ. Các phân tiếp theo sẽ mô tả cho bạn dùng SPSS như thế nào để đưa ra hầu hết các dạng thông thường của tóm tắt thống kê. Chúng được trình bày theo cách xác định kế hoạch phân tích được liệt kê ở trên cho các câu hỏi mô tả bộ số liệu mẫu.

3.5. Phân tích mô tả cho một biến


3.5.1. Một biến phân loại

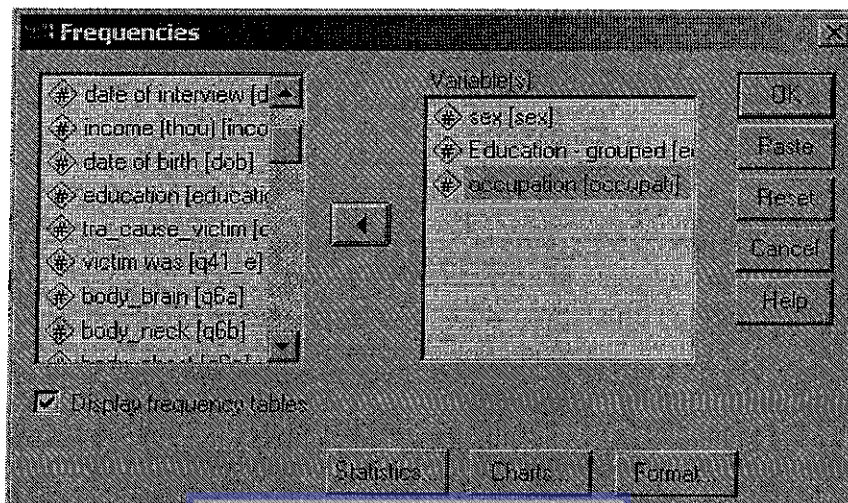
3.5.1.1. Bảng tần số

Các bảng tần số, biểu đồ cột tần suất và tỷ lệ là mô tả phân bố các giá trị của một biến phân loại. Bạn có thể dùng SPSS bằng các lệnh sau để tính các tần số, biểu đồ cột và tỷ lệ.



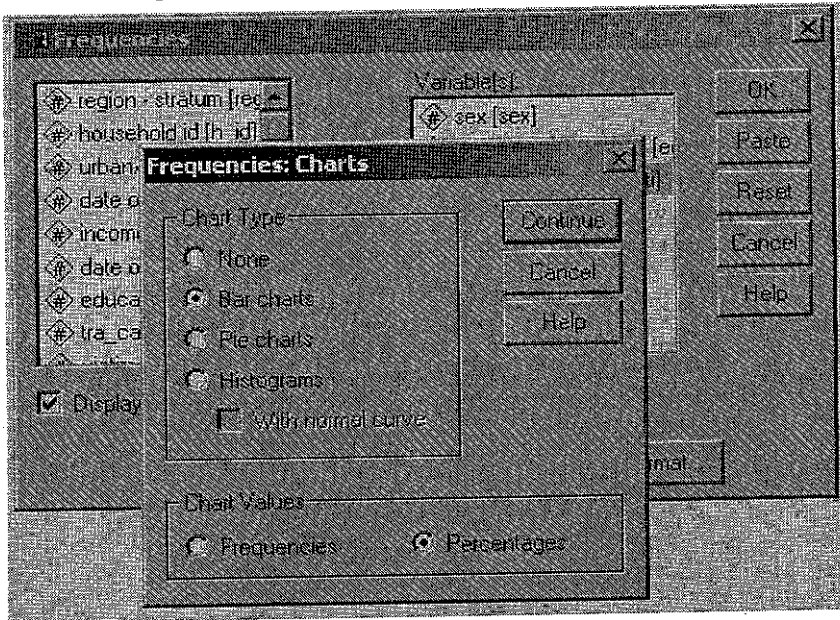
DÙNG SPSS ĐỂ TẠO MỘT BẢNG PHÂN BỐ TẦN SỐ

1. Từ thực đơn dọc chọn: Analyze /Descriptive Statistics/Frequencies.
2. Từ danh sách các biến, chọn biến sex (giới tính), edgrp (trình độ học vấn), và occupati (nghề nghiệp) và chuyển chúng vào hộp Variable(s) bằng cách nhấp chuột lên phím mũi tên 



3.5.1.2. Biểu đồ cột

Để vẽ biểu đồ cột biểu diễn bảng tần số, nhấp chuột lên Charts để. Bạn có thể chọn biểu đồ cột biểu diễn tần suất nhưng tốt hơn là biểu diễn dưới dạng các tỷ lệ (các tần số liên quan).



SPSS sẽ cho kết quả trong cửa sổ kết quả như sau:

Frequency Table

sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	1107	64.3	64.3	64.3
	female	614	35.7	35.7	100.0
	Total	1721	100.0	100.0	

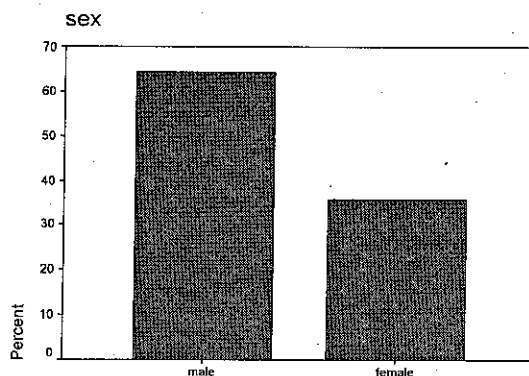
Education - grouped

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Children	83	4.8	4.9	4.9
	Less than secondary	474	27.5	27.8	32.7
	Secondary	991	57.6	58.2	91.0
	More than secondary	154	8.9	9.0	100.0
	Total	1702	98.9	100.0	
Missing	-1.00	19	1.1		
Total		1721	100.0		

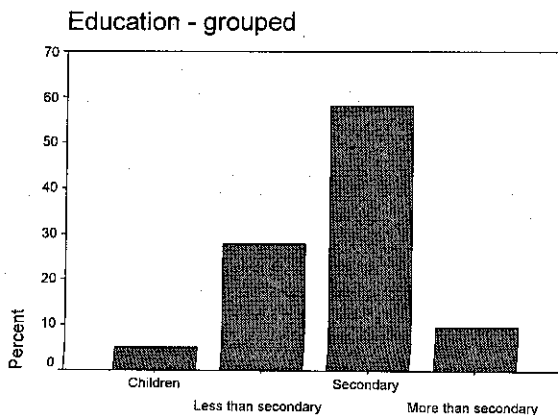
occupation

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	farmer	546	31.7	31.9	31.9
	gov. off	217	12.6	12.7	44.6
	petty tr	84	4.9	4.9	49.6
	employer	39	2.3	2.3	51.8
	studying	309	18.0	18.1	69.9
	children	129	7.5	7.5	77.5
	handicra	190	11.0	11.1	88.6
	retired	69	4.0	4.0	92.6
	poor hea	22	1.3	1.3	93.9
	unemploy	66	3.8	3.9	97.8
	other	38	2.2	2.2	100.0
	Total		1709	99.3	100.0
Missing	System	12	.7		
Total		1721	100.0		

Bar Chart

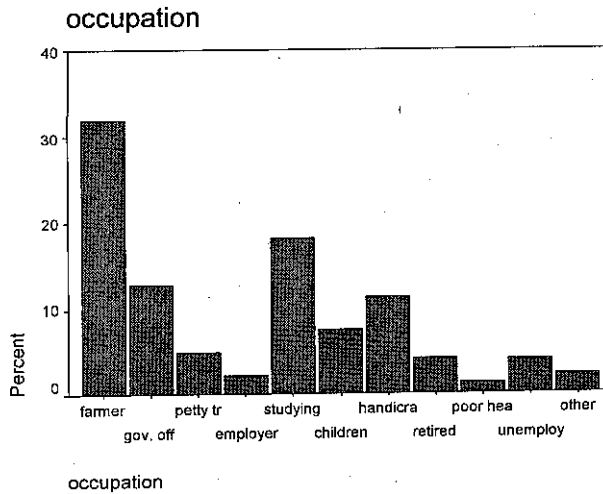


sex



Education - grouped





3.5.2. Một biến liên tục

Dùng bảng tần số cho một biến liên tục là không phù hợp, vì có rất nhiều giá trị được đưa vào bảng. Các tóm tắt thống kê bằng số, ví dụ các giá trị trung bình và các giá trị trung vị tương ứng lượng giá độ phân tán sẽ phù hợp hơn. Vì biến liên tục sử dụng kiểu đo lường tỷ số nên biểu đồ cột liên phù hợp hơn là biểu đồ cột. Trong bộ số liệu mẫu, chúng ta muốn mô tả phân bố của tuổi trong mẫu nghiên cứu.


Vì bạn chưa đánh giá được phân bố tuổi là phân bố chuẩn hay không? (xem mục 4.7). Vì thế bạn thực sự cần phải tính cả hai giá trị trung bình và trung vị, nhưng nếu bạn biết được phân bố tuổi là chuẩn hay không thì bạn mới có thể quyết định được là giá trị trung bình hay giá trị trung vị là thích hợp!

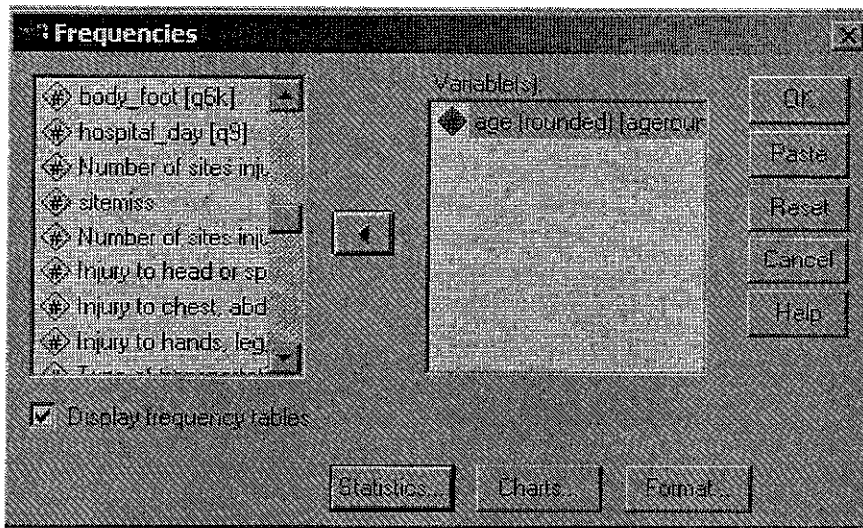
3.5.2.1. Trung bình và sự phân tán: Giá trị trung bình và độ lệch chuẩn

Giả sử rằng biến tuổi là phân bố chuẩn, và bạn muốn sử dụng giá trị trung bình và độ lệch chuẩn để mô tả phân bố của biến này.

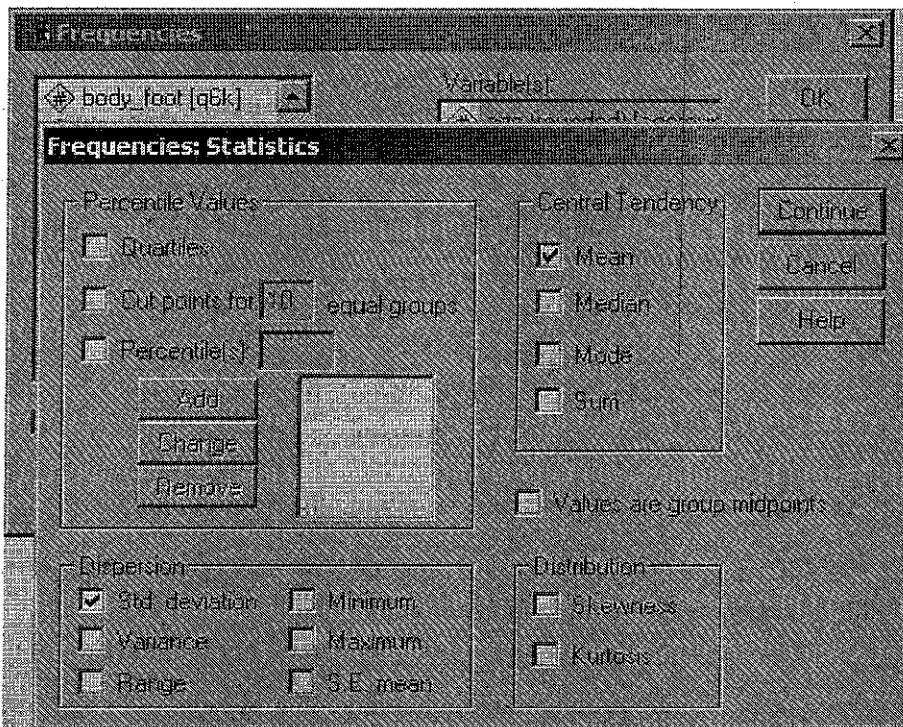


DÙNG SPSS ĐỂ TÍNH GIÁ TRỊ TRUNG BÌNH VÀ ĐỘ LỆCH CHUẨN

1. Từ thực đơn dọc chọn Analyse/Descriptive Statistics/Frequencies.
 2. Từ danh sách biến, chọn biến age (tuổi) và chuyển vào hộp Variable(s)
- bằng cách nhấp chuột lên biểu tượng 



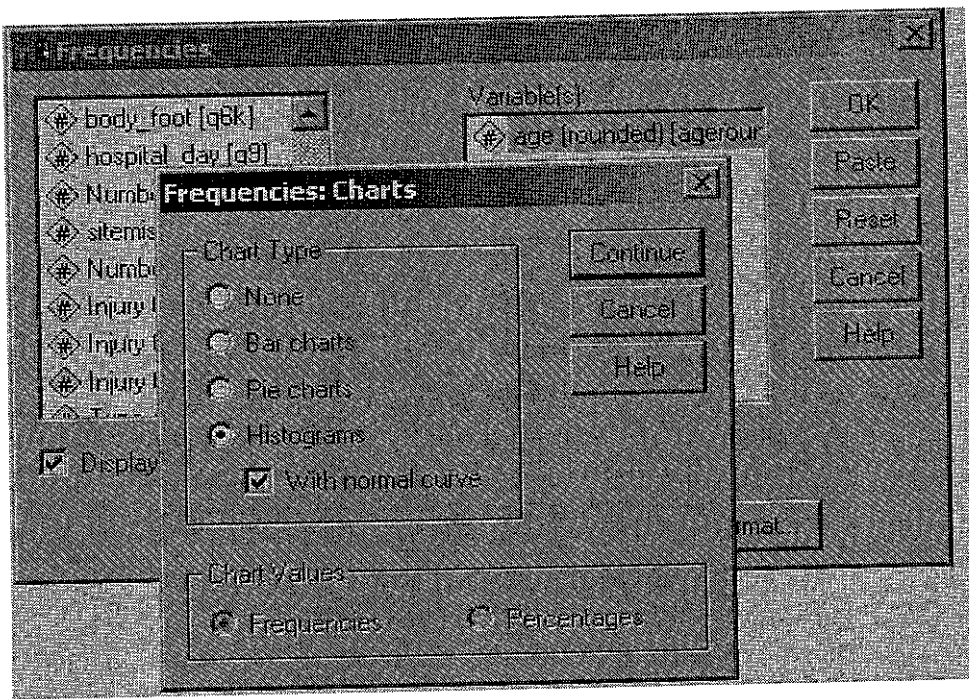
3. Nhấp chuột lên Statistics, chọn Mean và Std deviation.



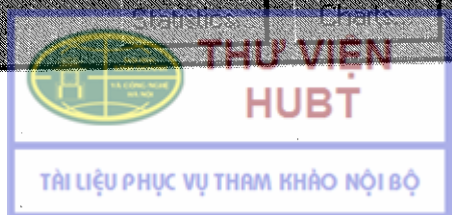
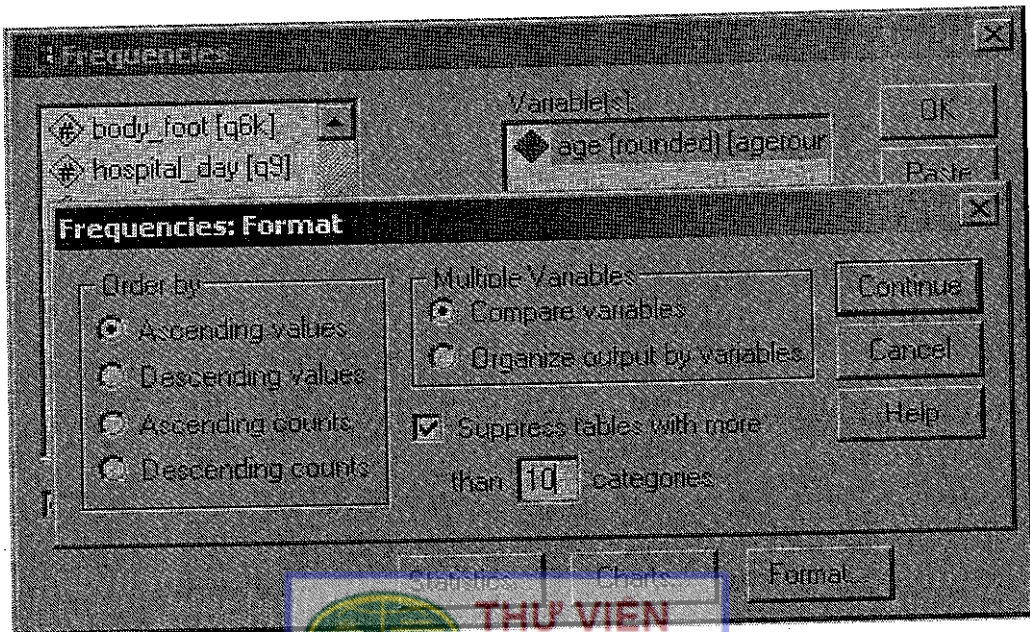
3.5.2.2. Biểu đồ cột liên

Khi tiến hành các bước tính trên bạn có thể yêu cầu vẽ biểu đồ biểu diễn sự phân bố ví dụ biểu đồ cột liên.

4. Nhấp chuột lên Continue trở về màn hình chính, nhấp chuột lên Charts, chọn Histogram và Normal curve.



5. Nhấp chuột lên Continue trở về thực đơn chính. Vì bạn chúng ta không muốn có bảng phân bố tần số mà có quá nhiều loại tuổi nên nhấp chuột lên nút Format, và trong thực đơn chọn Suppress... để phân loại cho biến.

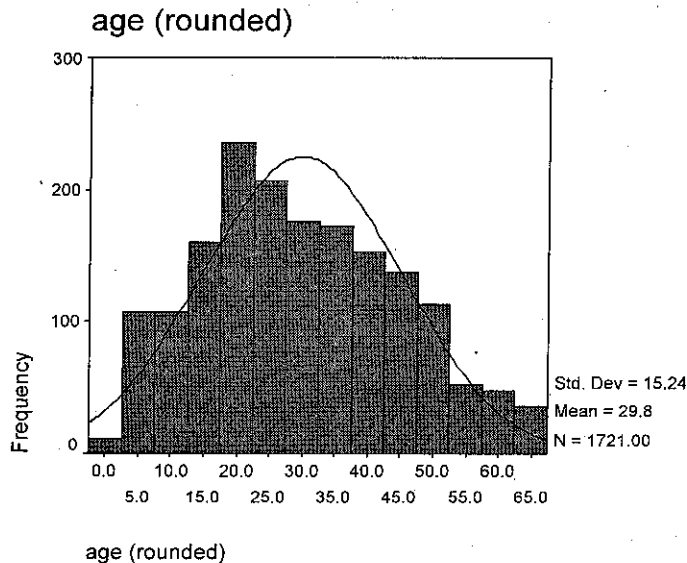


6. Nhấp chuột lên Continue để trở về màn hình chính, nhấp chuột lên OK để kết thúc lệnh. Kết quả sẽ xuất hiện trong màn hình kết quả của SPSS như sau:

Frequencies

Statistics

age (rounded)		
N	Valid	1721
	Missing	0
Mean		29.83
Std. Deviation		15.245



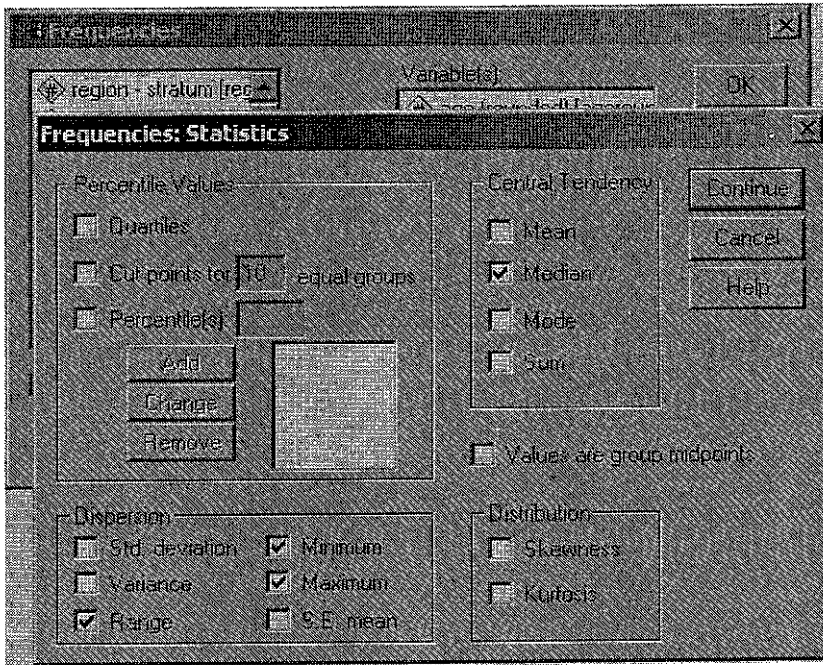
3.5.2.3. Trung bình và sự phân tán: Giá trị trung vị và khoảng

Giả sử biến tuổi không là phân bố chuẩn, và bạn cần có giá trị trung vị và một số giá trị lượng giá độ phân tán để mô tả phân bố của biến này. Đo lường độ phân tán bao gồm những giá trị đơn giản là giá trị cực tiểu và giá trị cực đại của phân bố, hoặc chính thức hơn là khoảng (cực tiểu – cực đại) hoặc lựa chọn một trong những khoảng phân vị, thường là phân vị thứ 2.5 và phân vị thứ 97.5 chứa 95% phân bố. Để tính toán các giá trị thống kê này bạn thực hiện các bước từ 1 đến 4 trong phần 3.5.2.1 ở trên sau đó thực hiện tiếp:

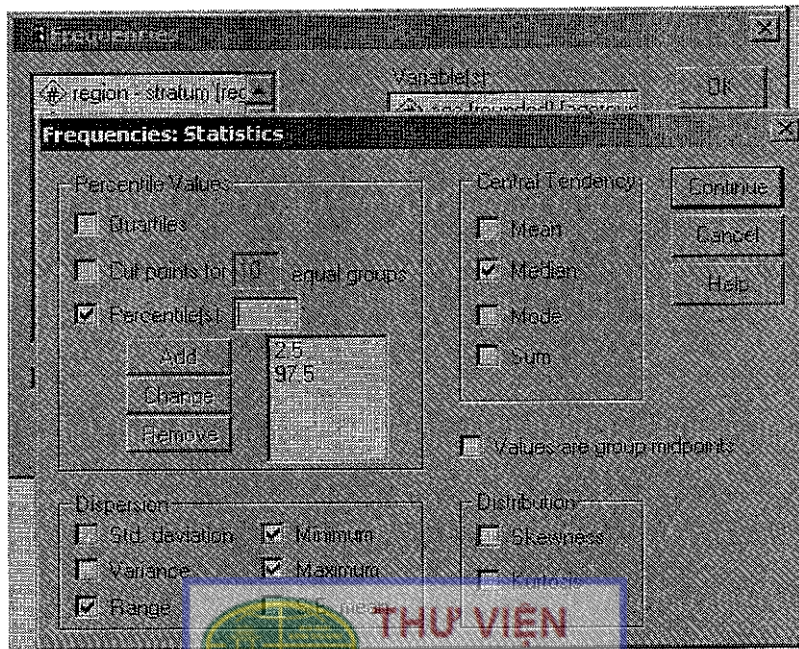


DÙNG SPSS ĐỂ TÍNH GIÁ TRỊ TRUNG VỊ VÀ ĐỘ PHÂN TÁN THỐNG KÊ

1. Nhấp chuột lên Statistics, chọn Median, Minimum, Maximum, và Range.



2. Nếu bạn muốn chọn phân vị thứ 2.5 và phân vị thứ 97.5, chọn hộp Percentiles, nhập số 2.5 vào hộp liên kế, nhấp chuột lên Add, nhập số 97.5 vào hộp này và nhấp chuột lên Add.



3. Nhấp chuột lên Continue/OK để hoàn thành lệnh. Phần kết quả thống kê của SPSS đưa ra như sau:

Statistics

age (rounded)

N	Valid	1721
	Missing	0
Median		28.00
Range		65
Minimum		0
Maximum		65
Percentiles	2.5	4.00
	97.5	61.95

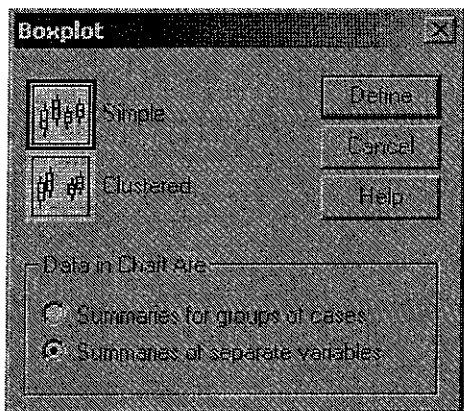
3.5.2.4. Biểu đồ Box-and-Whisker


Một dạng khác của biểu đồ biểu diễn biến liên tục là biểu đồ Box-and-Whisker, biểu đồ này phù hợp với cả phân bố chuẩn và không phải phân bố chuẩn.



DÙNG SPSS ĐỂ VẼ BIỂU ĐỒ BOX-AND-WHISKER

1. Từ thực đơn dọc chọn: **Graph/Boxplot**, chọn Summaries of Separate Variables/Define

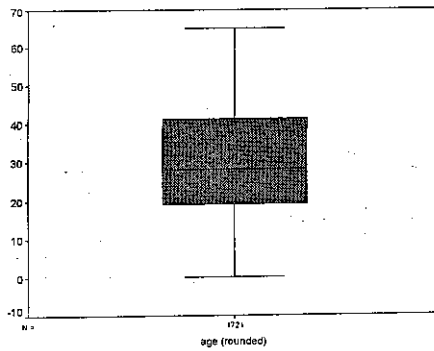


2. Từ danh sách các biến trong màn hình tiếp theo, chọn biến age (tuổi) và chuyển vào hộp Variable(s) bằng cách nhấp chuột lên 

3. Nhấp chuột lên OK để kết thúc lệnh.

Kết quả trong SPSS có dạng:





3.6. Mô tả các mối liên quan


3.6.1. Mô tả mối liên quan giữa biến phân loại với biến phân loại

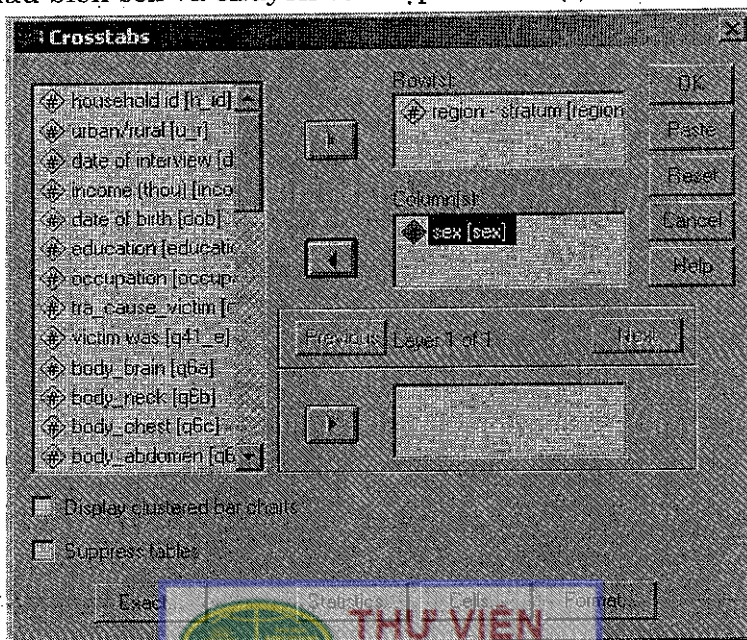
Mô tả về giới tính của mẫu nghiên cứu theo vùng.

Giới tính và vùng đều là các biến phân loại, vì thế kết quả SPSS CROSSTABS sẽ đưa ra kết quả cần thiết.

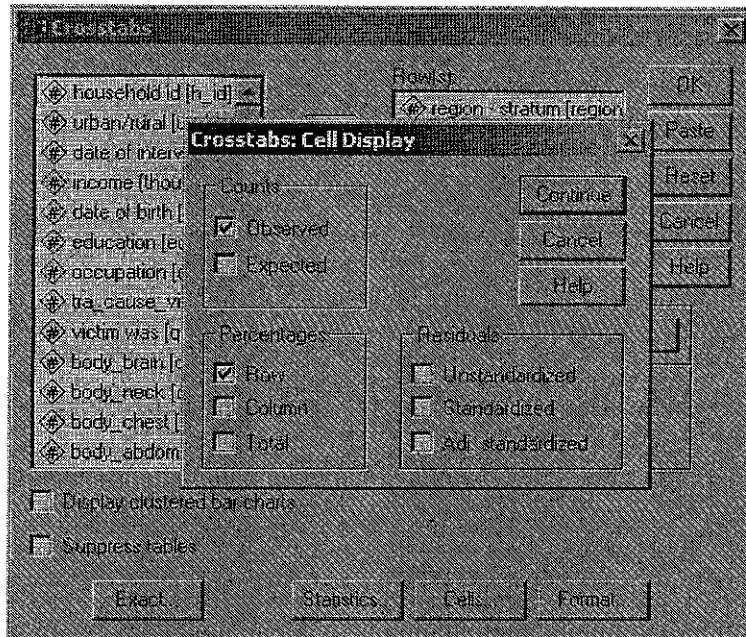


DÙNG SPSS TÍNH GIÁ TRỊ TRUNG BÌNH VÀ ĐỘ LỆCH CHUẨN

1. Từ thực đơn dọc chọn: Analyse/Descriptive Statistics /Crosstabs.
2. Từ danh sách biến, chọn biến region và chuyển vào hộp Row(s) bằng cách nhấp chuột lên .
3. Đánh dấu biến sex và chuyển vào hộp Column(s).



4. Nhấp chuột lên nút Cells và chọn nút Row để yêu cầu đưa kết quả tỷ lệ theo hàng (giới tính theo vùng).



5. Nhấp chuột lên Continue sau đó OK để hoàn thành lệnh.

Kết quả trong SPSS có dạng:

region - stratum * sex Crosstabulation

			sex		Total
			male	female	
region - stratum	ne	Count	141	59	200
		% within region - stratum	70.5%	29.5%	100.0%
	nw	Count	34	22	56
		% within region - stratum	60.7%	39.3%	100.0%
	rr	Count	124	76	200
		% within region - stratum	62.0%	38.0%	100.0%
	nc	Count	171	88	259
		% within region - stratum	66.0%	34.0%	100.0%
	cc	Count	96	56	152
		% within region - stratum	63.2%	36.8%	100.0%
	ch	Count	107	47	154
		% within region - stratum	69.5%	30.5%	100.0%
	se	Count	194	99	293
		% within region - stratum	66.2%	33.8%	100.0%
	mr	Count	240	167	407
		% within region - stratum	59.0%	41.0%	100.0%
Total		Count	1107	614	1721
		% within region - stratum	64.3%	35.7%	100.0%

Bảng này không phù hợp để trong báo cáo, vì dạng bảng không quen thuộc với người đọc. Tuy nhiên, trong bảng này có toàn bộ các thông tin để cung cấp cho các bảng khác để đọc hơn trong kế hoạch nghiên cứu.

3.6.2. Mối liên quan giữa một biến liên tục và một biến phân loại

Mối liên quan giữa một biến liên tục và một biến phân loại có thể được mô tả theo cả hai cách bằng số và biểu đồ.

Hãy cân nhắc hai cách tiếp cận cho câu hỏi nghiên cứu thứ hai,


Mô tả sơ lược tuổi của mẫu nghiên cứu theo vùng

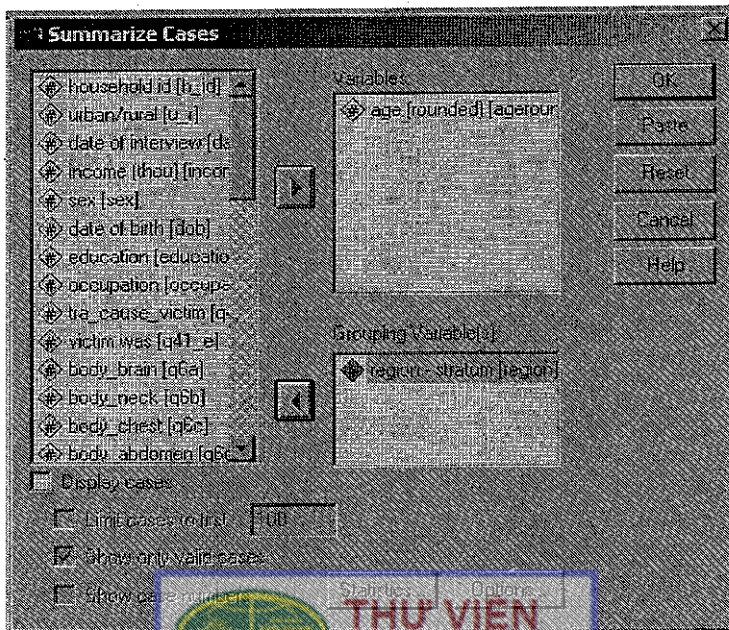
3.6.2.1. Mô tả bằng số

Tuổi là biến liên tục và vùng là biến phân loại. Có nhiều cách để có giá trị trung bình (trung vị) trong các nhóm của biến khác. Cách toàn diện nhất là:

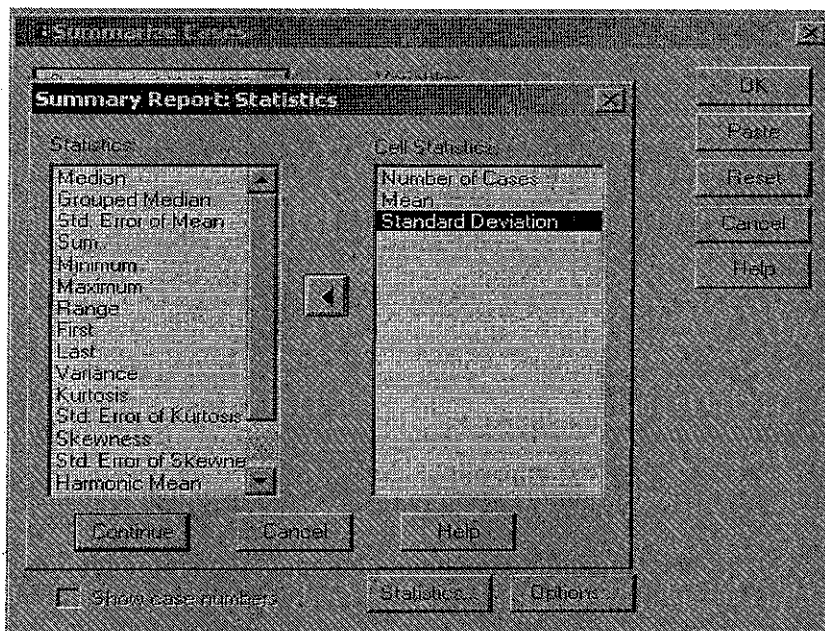


DÙNG SPSS ĐỂ TÍNH GIÁ TRỊ TRUNG BÌNH TRONG TỪNG PHÂN NHÓM

1. Từ thực đơn chọn: Analyze→Reports→Case Summaries
2. Từ danh sách biến, chọn biến *ageround* (tuổi) và chuyển vào hộp biến, sau đó chọn *region* và chuyển vào hộp Grouping Variable(s) bằng cách nhấp chuột lên 
4. Bỏ đánh dấu ở *Display Cases* – bạn không muốn điều này.



5. Nhấp chuột lên nút Statistics, và bôi đen Mean và Standard Deviation, sau đó chuyển chúng qua hộp Cell Statistics.



6. Nhấp chuột lên nút Continue, sau đó là nút OK để hoàn thành lệnh. Kết quả trong SPSS có dạng như hình dưới đây:

Case Summaries

age (rounded)

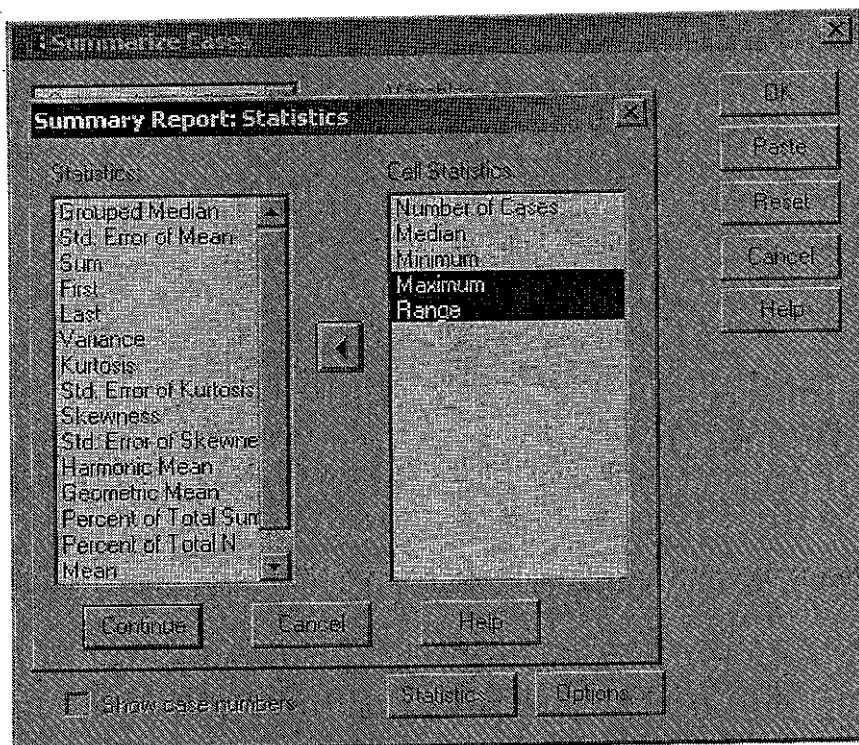
region - stratum	N	Mean	Std. Deviation
ne	200	31.47	13.066
nw	56	32.43	13.042
rr	200	31.88	15.689
nc	259	30.41	16.515
cc	152	29.32	12.943
ch	154	26.82	14.535
se	293	29.47	14.820
mr	407	28.89	16.582
Total	1721	29.83	15.245



DÙNG SPSS TÍNH GIÁ TRỊ TRUNG VỊ TRONG CÁC PHÂN NHÓM

Để có được giá trị trung vị thay vì giá trị trung bình, lặp lại bước 1 đến 3 ở trên, sau đó làm tiếp bước 4 dưới đây:

4. Nhấp chuột lên nút Statistics, bôi đen Median, Minimum, Maximum, và Range, rồi chuyển vào hộp Cell Statistics.



5. Nhấp chuột lên nút Continue, sau đó chọn OK để hoàn thành lệnh.

Kết quả trong SPSS có dạng như dưới đây

Case Summaries

age (rounded)

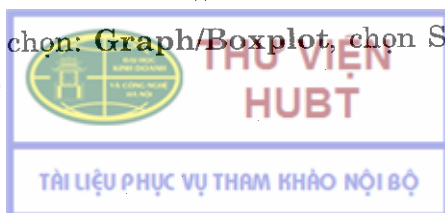
region - stratum	N	Median	Minimum	Maximum	Range
ne	200	31.00	0	65	65
nw	56	32.00	5	59	54
rr	200	31.00	3	64	61
nc	259	29.00	2	65	63
cc	152	28.00	4	61	57
ch	154	26.00	2	65	63
se	293	27.00	1	63	62
mr	407	24.00	0	65	65
Total	1721	28.00	0	65	65

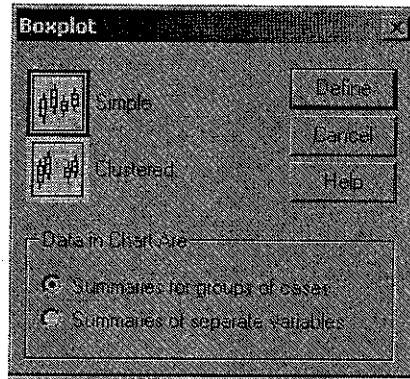
3.6.2.2. Tóm tắt bằng biểu đồ Boxplot trong các phân nhóm



DÙNG SPSS VẼ BIỂU ĐỒ BOXPLOTS TRONG CÁC PHÂN NHÓM

1. Từ thực đơn dọc chọn: Graph/Boxplot, chọn Simple, và Summaries for Groups of Cases/Define

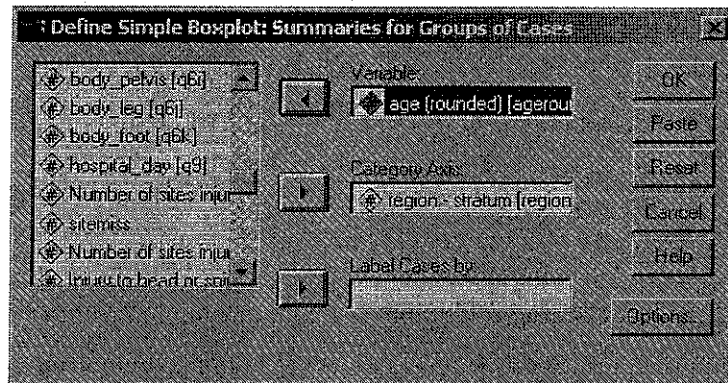




2. Từ danh sách các biến trong màn hình tiếp theo, chọn biến ageround (tuổi) và chuyển vào trong hộp Boxes Represent bằng cách nhấp chuột lên

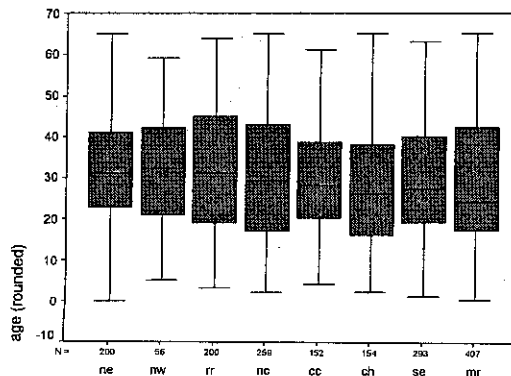


3. Từ danh sách biến trong màn hình tiếp theo, chọn biến region và chuyển vào hộp Category Axis bằng cách nhấp chuột lên



4. Nhấp chuột lên OK để thực hiện lệnh.

Kết quả trong SPSS có dạng sau:



3.6.3. Mối liên quan giữa một biến liên tục với một biến liên tục

Mối liên quan giữa hai biến liên tục có thể được tóm tắt dưới dạng số như là hệ số tương quan hoặc biểu đồ chấm. Với câu hỏi nghiên cứu thứ 3 trong bộ số liệu mẫu:


Mô tả chất lượng cuộc sống trước chấn thương, và xem có sự ảnh hưởng của tuổi đến chất lượng cuộc sống trước chấn thương hay không.

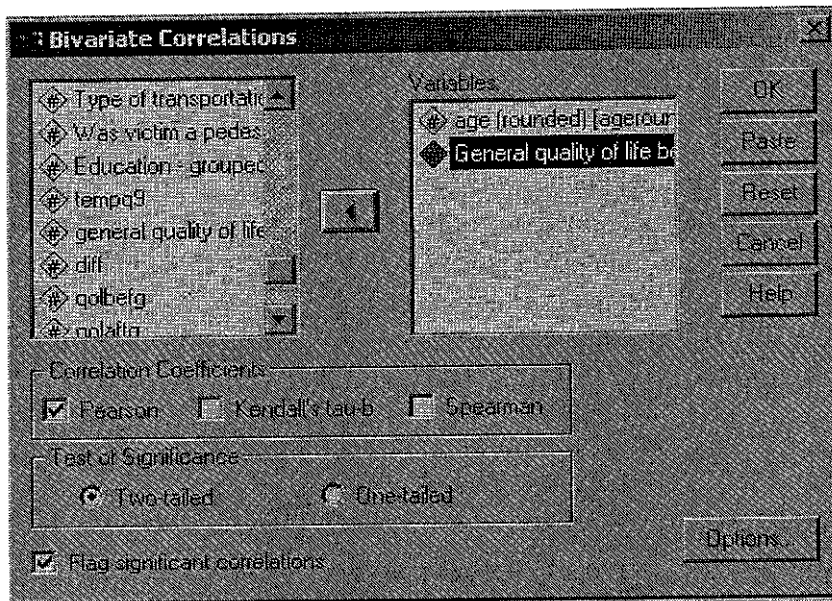
Tuổi là một biến liên tục và điểm chất lượng cuộc sống cũng là biến liên tục. Nếu cả hai biến đều có phân bố chuẩn thì dùng hệ số tương quan Pearson, nếu không phải sử dụng hệ số tương quan Spearman.

3.6.3.1. Tóm tắt bằng số – Các hệ số tương quan



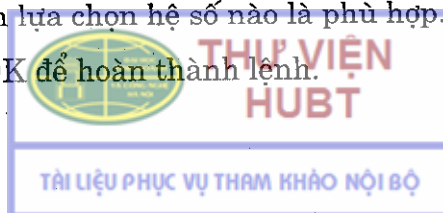
DÙNG SPSS ĐỂ TÍNH HỆ SỐ TƯƠNG QUAN

1. Từ thực đơn chọn: Analyse/Correlate/Bivariate.
2. Từ danh sách biến, bôi đen biến *ageround* (tuổi) và *region* sau đó chuyển vào hộp biến bằng cách nhấp chuột lên 



3. Hệ số Pearson được mặc định trong SPSS, vì thế không cần thay đổi trừ khi bạn muốn tính hệ số Spearman. Không thể yêu cầu cả hai – kế hoạch phân tích của bạn sẽ giúp bạn lựa chọn hệ số nào là phù hợp.

4. Nhấp chuột lên OK để hoàn thành lệnh.



Kết quả trong SPSS cho hệ số tương quan Pearson's có dạng sau:

Correlations

		age (rounded)	General quality of life before injury
age (rounded)	Pearson Correlation	1	.059*
	Sig. (2-tailed)		.015
	N	1721	1692
General quality of life before injury	Pearson Correlation	.059*	1
	Sig. (2-tailed)	.015	
	N	1692	1692

*. Correlation is significant at the 0.05 level (2-tailed).

Yêu cầu hệ số Spearman's là kết quả của bảng tiếp theo:

Correlations

			age (rounded)	General quality of life before injury
Spearman's rho	age (rounded)	Correlation Coefficient	1.000	.065**
		Sig. (2-tailed)		.007
		N	1721	1692
	General quality of life before injury	Correlation Coefficient	.065**	1.000
		Sig. (2-tailed)	.007	
		N	1692	1692

** Correlation is significant at the .01 level (2-tailed).

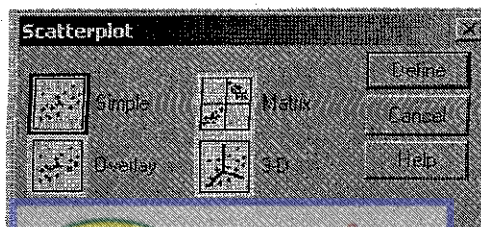
3.6.3.2. Mô tả bằng biểu đồ – biểu đồ chấm điểm

Để yêu cầu vẽ biểu đồ chấm về mối liên quan giữa hai biến liên tục, đầu tiên bạn cần định rõ biến nào nằm trên trục X và biến nào sẽ nằm trên trục Y. Trong ví dụ này, chúng ta thấy chất lượng cuộc sống chịu ảnh hưởng của tuổi hợp lý hơn là tuổi chịu ảnh hưởng của chất lượng cuộc sống, vì thế tuổi sẽ nằm trên trục X và chất lượng cuộc sống nằm trên trục Y.



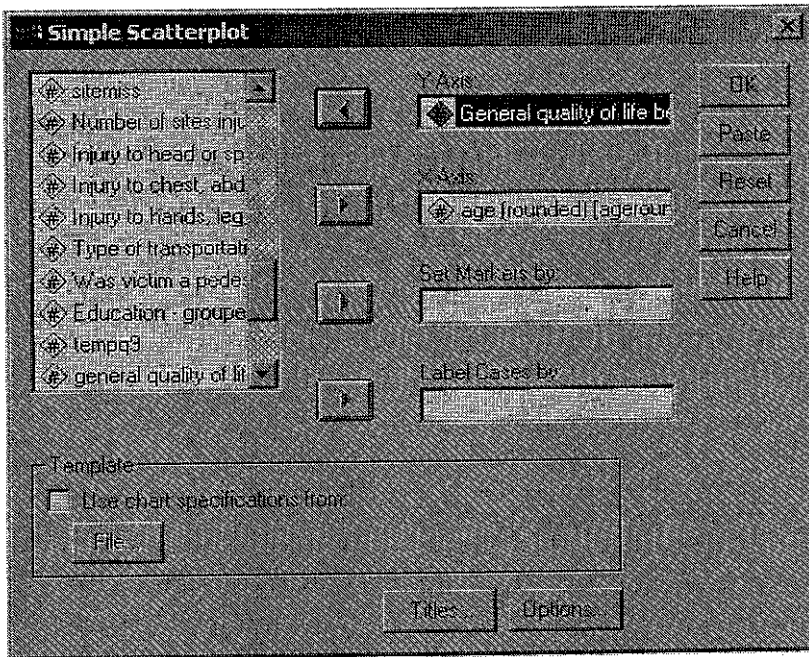
DÙNG SPSS VẼ BIỂU ĐỒ CHẤM ĐIỂM

1. Từ thực đơn chọn: Graph/Scatter/Simple

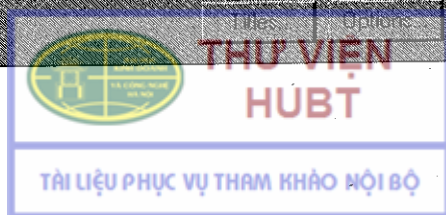
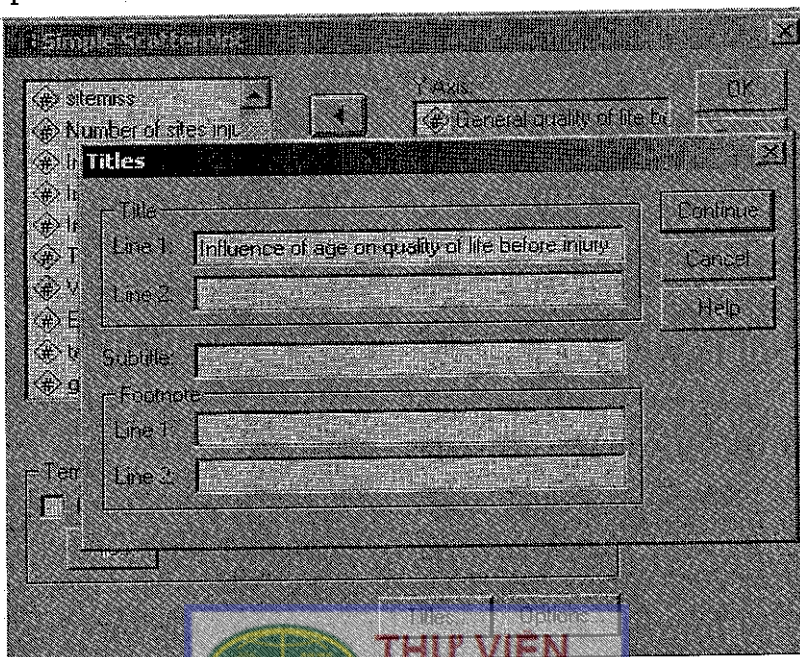


2. Nhấp chuột lên Define, sau đó từ danh sách biến, chọn biến ageround (tuổi) và chuyển vào hộp X-axis, sau đó chọn biến qol_bef (Điểm chất lượng cuộc sống trước chấn thương) và chuyển vào trong hộp Y-axis bằng cách nhấp chuột

lên 

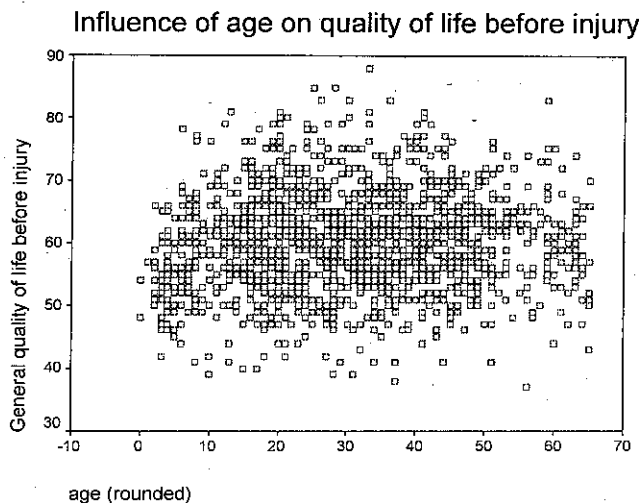


3. Nhấp chuột lên Titles và kết quả một tiêu đề phù hợp sẽ xuất hiện trong biểu đồ kết quả.



4. Nhấp chuột lên Continue, sau đó là OK để hoàn thành lệnh.

Kết quả trong SPSS có dạng sau:



3.7. Viết kết quả của phân tích mô tả

Một trong những phần tốt nhất của thực hiện kế hoạch phân tích trên trong phân tích của bạn là một số phần của báo cáo cũng được viết. Bạn sẽ có đủ thông tin để viết phần phương pháp nghiên cứu, và có ý kiến về phần kết quả sẽ có dạng như thế nào.

Trong ví dụ mẫu, phần tiếp theo đây tóm tắt từ kế hoạch phân tích và các kết quả đạt được từ những phân tích mô tả ở trên. Đây là một ví dụ gợi ý cho bạn cách viết kết quả từ các nghiên cứu tương tự, nhưng mỗi người đều có phong cách riêng của mình.

Báo cáo mẫu

Phương pháp thống kê

Thống kê mô tả được dùng cho 3 mục tiêu nghiên cứu đầu tiên, những mục tiêu mô tả sơ lược các yếu tố xã hội – dân số của các đối tượng chấn thương giao thông năm 2001 trong mẫu nghiên cứu, để xác định sự khác nhau về phân bố giới tính và tuổi giữa các vùng và xác minh có hay không mối liên quan giữa tuổi và chất lượng cuộc sống trước chấn thương.

Vì các biến liên quan có các loại khác nhau, một loạt các phân tích thống kê đã được thực hiện. Phân bố của tuổi được xác định là phân bố chuẩn nên giá trị trung bình và độ lệch chuẩn được chọn để tóm tắt biến này trong toàn bộ quần thể cho mục tiêu đầu tiên và cho từng vùng miền cho mục tiêu thứ hai. Tần suất và tỷ lệ được dùng để tóm tắt biến giới tính, trình độ học vấn, nghề nghiệp cho



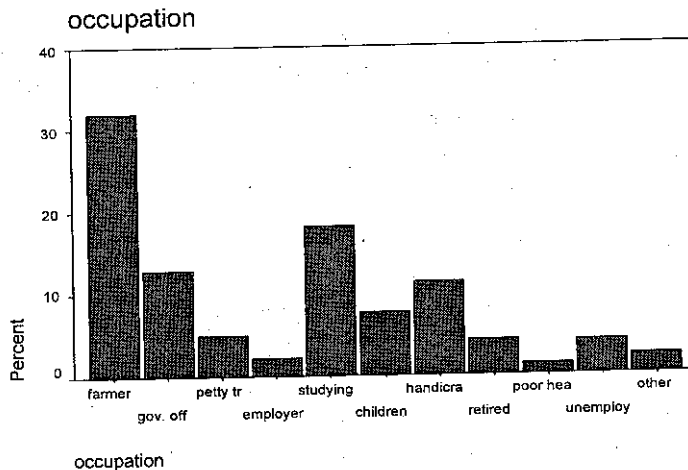
mục tiêu đầu, và mô tả sự khác nhau về tuổi giữa các vùng cho mục tiêu thứ hai. Vì điểm chất lượng cuộc sống cũng được xác định là phân bố chuẩn nên hệ số tương quan Pearsons được dùng để tóm tắt mối liên quan giữa tuổi và chất lượng cuộc sống cho mục tiêu thứ ba.

Lưu ý rằng phần phương pháp nên viết càng ngắn càng tốt, và tránh sự nhắc lại khi bạn sử dụng các tóm tắt và kiểm định giống nhau trong nhiều lần. Những lý do giải thích tại sao lại chọn kiểm định này cũng cần phải trình bày rõ ràng (vì là phân bố chuẩn hay phân loại...).

Kết quả

Các đặc điểm xã hội – dân số:

Mẫu nghiên cứu bao gồm 1721 đối tượng chấn thương giao thông năm 2001 có tuổi từ 65 trở xuống. Gần 2/3 (1107, 64%) đối tượng là nam, tuổi trung bình là 29,8 tuổi (độ lệch chuẩn, $sd = 15,2$ tuổi). Hơn một nửa có trình độ học vấn là THCS (991, 58%), 1/3 đối tượng có TĐHV dưới THCS (5% là trẻ em chưa đến tuổi đi học), và 9% còn lại có trình độ học vấn sau THSC. Về nghề nghiệp, khoảng 1/3 (546) đối tượng làm ruộng, 1/5 còn đang đi học, và trên 10% đối tượng là CBCNV hoặc công nhân thủ công. Nhiều nhóm nghề nghiệp khác có tỷ lệ nhỏ, số liệu được trình bày trong biểu đồ 3.1.



Biểu đồ 3.1. Phân bố nghề nghiệp của 1721 đối tượng chấn thương giao thông dưới 65 tuổi năm 2001

Tiêu đề đã được sử dụng phản ánh mục tiêu nghiên cứu đầu tiên. Lưu ý rằng chỉ có một biến được mô tả trong biểu đồ này. Giữ các biểu đồ mô tả không dễ dàng viết bằng các từ. Nghề nghiệp là một biến phức tạp, vì thế chỉ một số nghề được biểu thị bằng từ ngữ và những người đọc được cho phép tự tìm hiểu. Trình độ học vấn, với 4 phân loại thì dễ mô tả bằng từ ngữ hơn.

Những sự khác nhau về dân số giữa các vùng.

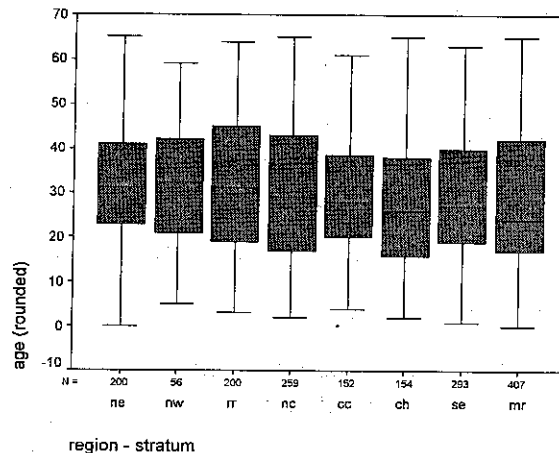
Khoảng 2/3 đối tượng nghiên cứu là nam giới. Tỷ lệ này chạy từ 59% ở Vùng châu thổ sông Mêkông lên 70% ở vùng Tây Bắc.

Bảng phân bố giới tính theo từng vùng.

Bảng.... Phân bố giới tính theo vùng của nghiên cứu

Vùng	Số người	Tỷ lệ Nam (%)
Tây Bắc	200	70,5
Đông Bắc	56	60,7
ĐB Sông Hồng	200	62,0
Bắc Trung bộ	259	66,2
Nam Trung bộ	152	63,2
Tây Nguyên	154	69,5
Đông Nam Bộ	293	66,2
ĐB Sông Mê Kông	407	59,0
Chung	1721	64,3

Tuổi trung bình của mẫu nghiên cứu là 30 và tương tự ở hầu hết các vùng, mặc dù các đối tượng ở Tây Nguyên độ tuổi trung bình trẻ hơn (27 tuổi) thì sự phân bố tuổi ở tất cả các vùng là giống nhau.



Biểu đồ 3.2. Phân bố tuổi theo vùng

Tiếp theo phần này, sự pha trộn các thông tin của bảng và biểu đồ đã được viết, mặc dù cả hai kết quả có thể được biểu diễn dưới dạng bảng (như kế hoạch

ban đầu). Biểu đồ Boxplot cung cấp một bức tranh rõ ràng hơn về sự phân bố của tuổi trong từng vùng, so với kết quả được giới thiệu sẵn trong bảng kết quả bằng độ lệch chuẩn (bạn có thể không đồng ý).

Chất lượng cuộc sống trước chấn thương.

Điểm trung bình QoL trước chấn thương là 60.4 (độ lệch chuẩn là 7.7). Mối tương quan giữa điểm QoL trước chấn thương và tuổi là không đáng kể (Pearson's $r = 0.06$).

Điểm trung bình QoL ở trên được lấy từ những chỉ dẫn trong phần 3.5.2, nhưng nó không được trình bày chi tiết trong bài này. Ở đây, kết quả SPSS cho thấy hệ số tương quan Pearson's = 0,06 (quá bé so với 1), cho nên bạn có thể kết luận là không có mối liên quan giữa tuổi và điểm chất lượng cuộc sống trước chấn thương. Trong viết báo cáo bạn nên dừng lại ở kết luận này mà không cần thiết vẽ biểu đồ biểu diễn nữa.

Bài số

4

PHÂN TÍCH THỐNG KÊ SUY LUẬN

Mục tiêu

5. Chọn được kiểm định thống kê phù hợp để so sánh các kiểu dữ liệu khác nhau.
6. Trình bày được lý do cơ bản của việc dùng các phép tính toán của mỗi kiểm định thống kê.
7. Trình bày được các giả định liên quan đến mỗi kiểm định thống kê.
8. Sử dụng được SPSS để thực hiện các kiểm định thống kê.
9. Phiên giải được các kết quả phân tích số liệu cho mỗi loại kiểm định.

4.1. Giới thiệu

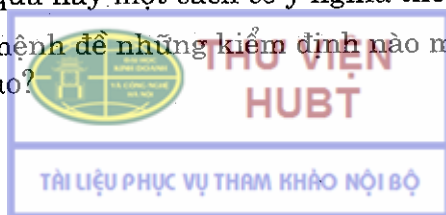
Bài này mô tả một cách cơ bản nhất các kiểm định thống kê thường hay được sử dụng. Mục đích của khóa học Thống kê y tế công cộng này là giúp bạn hiểu được làm thế nào để chọn được kiểm định thống kê đúng với các điều kiện của thiết kế nghiên cứu, phiên giải và viết báo cáo kết quả tính toán của kiểm định thống kê. Khóa học này không hướng dẫn bạn tính toán tất cả các kiểm định thống kê, nó chỉ cung cấp cho bạn các khái niệm chung về cách làm thế nào để phân tích số liệu, cách để chọn đúng loại kiểm định thống kê, cách để kiểm tra tính xác thực của các kết quả phân tích đó, cách sử dụng SPSS để phân tích số liệu và cách để phiên giải các kết quả đầu ra của kiểm định thống kê. Bất cứ một quyển sách thống kê “tốt” nào cũng có thể cung cấp cho bạn các công thức tính nếu bạn quan tâm đến nó mặc dù các phép tính này đã được thực hiện bởi các phần mềm phân tích số liệu. Trong bài này, chúng tôi cũng cung cấp cho bạn một bảng liệt kê các kiểm định thống kê và các tình huống khác nhau để giúp bạn chọn lựa. Nếu thiết kế nghiên cứu và cách thu thập số liệu của bạn không phù hợp với bất kỳ một cách nào trong bảng, bạn hãy tham khảo ý kiến của các chuyên gia thống kê.

4.2. Quá trình lập kế hoạch phân tích số liệu

Phân tích thống kê của một bộ số liệu sẽ không khó nếu số liệu đã được làm sạch và chuẩn bị tốt (xem bài 2), các giả thuyết nghiên cứu đã được định nghĩa rõ ràng (xem bài 1). Phần khó khăn của phân tích số liệu là định nghĩa các câu hỏi để phân tích, phần còn lại chỉ là một quá trình mang tính “công thức” và chúng ta có thể bắt chước. Giáo trình và khóa học này sẽ cung cấp cho các bạn “công thức” để tiến hành phân tích thống kê cơ bản cho các nghiên cứu.

Bạn nên chuẩn bị một kế hoạch làm thế nào để tóm tắt và phân tích số liệu. Sau đây là một số câu hỏi mà bạn cần phải trả lời để giúp bạn chuẩn bị kế hoạch:

1. Các giả thuyết khoa học nào có thể kiểm định được (giả thuyết thống kê) mà câu hỏi nghiên cứu đề cập đến. Một giả thuyết có thể kiểm định được bao gồm một mệnh đề cho cả giả thuyết không (H_0) và đối thuyết (H_1). Như bạn đã thấy, thông thường có nhiều hơn một giả thuyết thống kê từ một câu hỏi nghiên cứu.
2. Với mỗi giả thuyết thống kê:
 - (i) Liệt kê danh sách các biến
 1. Định nghĩa biến độc lập và biến phụ thuộc.
 2. Xác định các mức độ đo lường của các biến độc lập và biến phụ thuộc (liên tục /khoảng hay phân loại).
 - (ii) Mô tả các mối liên quan
 1. Sử dụng các thông tin từ (i), chọn cách mà bạn sẽ tóm tắt mối quan hệ giữa hai biến (bài 3).
 2. Chỉ ra các “bảng giả” mô tả mối quan hệ này trong báo cáo cuối cùng của bạn.
 - (iii) Chọn kiểm định thống kê
 1. Sử dụng bảng 1(tr.193) và bảng 2(tr.194), cùng với các thông tin từ cả (i) và (ii) chọn một loại kiểm định thống kê phù hợp nhất.
 2. Kiểm tra lại các giả định của kiểm định thống kê (xem mục 4.7).
 3. Sau đó đưa ra lựa chọn loại kiểm định thống kê nếu các giả định không thoả mãn.
 - (iv) Phiên giải các kết quả có ý nghĩa thống kê
 1. Định ra mức ý nghĩa thống kê mà bạn sẽ áp dụng để kiểm định giả thuyết.
 2. Viết ra những gì mà bạn muốn viết/nói trong báo cáo cuối cùng của bạn tới những người không phải là nhà thống kê (chỉ ra là bạn đã thực hiện và tìm thấy các kết quả này một cách có ý nghĩa thống kê).
 3. Hãy viết ra một mệnh đề những kiểm định nào mà bạn chọn để kiểm định thống kê và tại sao?



4.3. Giả thuyết thống kê

Việc chọn kiểm định thống kê để sử dụng khi so sánh dữ liệu được quyết định bởi loại biến và loại thiết kế nghiên cứu. Trong khi các phép tính toán cho mỗi kiểm định là khác nhau thì mối quan hệ giữa chúng là luôn luôn giống nhau. Số liệu được thu thập và được tính ra các đại lượng thống kê mô tả (trung bình, trung vị, tỷ lệ). Các đại lượng đó được so sánh để tóm tắt cho một bộ số liệu giả thuyết trong đó giả thuyết không được giả định là đúng. Từ thống kê mô tả và quan sát được, các đại lượng thống kê khác cũng được tính toán gọi là giá trị kiểm định (chúng ta ký hiệu là T). Đó là sự không thống nhất giữa các giá trị thống kê quan sát được và các giá trị kỳ vọng (ví dụ: sự khác nhau của trung bình quan sát và trung bình kỳ vọng), thông thường được chuẩn hóa thông qua một vài cách (chia cho một vài giá trị nào đó, chẳng hạn như là sai số chuẩn) để tạo ra một giá trị cuối cùng duy nhất. Các số liệu biến thiên, cũng như các kiểm định thống kê, phụ thuộc vào quá trình chọn mẫu. Thường thì chúng ta không đoán trước được cá nhân nào được chọn vào mẫu trên quần thể. Do đó chúng ta không thể biết được T nào là phù hợp nhưng những nhà thống kê đã thực hiện rất nhiều phép tính toán phức tạp để mang lại cho chúng ta một khoảng (khoảng điều kiện) của giá trị T có khả năng xảy ra nếu giả thuyết không đúng. Nếu T rơi vào ngoài cùng khoảng của chúng ta, thì chúng ta chỉ ra rằng giả thuyết không không đúng và kết luận rằng giả thuyết đối là hợp lý (mặc dù chúng ta không bao giờ đảm bảo được điều đó).

4.4. Sử dụng kiểm định nào?

Việc chọn cách để tóm tắt số liệu thống kê được chỉ định bởi thang đo của biến phụ thuộc, việc chọn kiểm định thống kê được quy định bởi thang đo của biến phụ thuộc và dạng so sánh. Trong giáo trình này có kèm theo hai bảng cách chọn loại kiểm định thống kê phù hợp (bảng 1 và 2 ở cuối sách), bảng sẽ tóm tắt cách chọn các kiểm định thống kê một cách đơn giản nhất cho biến phụ thuộc là liên tục và biến phân loại (biến đầu ra). Hãy xem nội dung của bảng ở trang 194, 195 và sử dụng các bảng đó để làm ví dụ sau đây.

Trong kiểm định thống kê, nếu bạn có một biến độc lập có phân bố chuẩn bạn có thể đi theo nhánh Có trong bảng 1. Các kiểm định đó được biết như là *kiểm định tham số (parametric test)*. Mục 4.7 sẽ miêu tả sẽ mô tả cho bạn biết một biến có dạng phân bố chuẩn hay không. Nếu biến phụ thuộc của bạn không có phân bố chuẩn thì bạn theo nhánh KHÔNG trong bảng, các kiểm định ở đây gọi là *kiểm định phi tham số (non-parametric test)*.

Chúng ta sẽ có cùng một kết quả bất kể chúng ta sử dụng kiểm định tham số hay phi tham số khi biến đầu ra có phân bố chuẩn, tuy nhiên kiểm định

tham số sẽ mạnh hơn trong việc đưa ra kết quả. Điều quan trọng nữa là chúng ta CHỈ sử dụng kiểm định phi tham số khi các giả định chuẩn không thỏa mãn.

Quay trở lại bảng 1(tr.193). Câu hỏi tiếp theo mà bạn sẽ gặp phải là loại so sánh mà bạn quan tâm trong giả thuyết. Đó là nguyên nhân tại sao phải đặt giả thuyết thống kê rõ ràng, và bạn muốn so sánh cái gì, tương ứng bạn cần xác định được loại biến độc lập là gì? Trong một vài trường hợp không có biến độc lập, và câu hỏi đơn giản là giá trị mẫu của chúng ta có quan hệ gì với giá trị giả thuyết khác trong quần thể hay không. Đôi khi biến độc lập có dạng liên tục và không có sự so sánh nhóm. Phổ biến nhất, giả thuyết thể hiện biến phụ thuộc khác biệt như thế nào giữa các nhóm (hai hay nhiều hơn) hoặc sự khác biệt theo thời gian trong cùng một nhóm. Loại giả thuyết thống kê chỉ ra cột nào trong bảng mà bạn quan tâm và điều này sẽ giúp bạn quyết định chọn loại kiểm định phù hợp.

Bạn có thể tìm thấy tất cả các loại kiểm định thống kê trong bảng 1(tr.193) và 2(tr.194) trong bất kỳ một cuốn sách thống kê cơ bản nào. Ngoài ra cuốn sách Thống kê Y tế Công cộng phần Thống kê cơ bản cũng đã đề cập đến các kiểm định này với đầy đủ cách tính và các ví dụ minh họa. Mục đích của khóa học này không yêu cầu các bạn nhớ lại cách tính các đại lượng thống kê nhưng bạn phải biết được cách để chọn loại kiểm định thống kê thích hợp, cách dùng SPSS để tính các kiểm định thống kê và làm thế nào để phiên giải kết quả thống kê. Phần lớn các phần mềm phân tích số liệu sẽ đưa ra các đại lượng thống kê cơ bản và tính toán cho các bạn giá trị kiểm định. Ngoài ra phần mềm cũng cho bạn biết giá trị p và ý nghĩa thống kê. Bạn sẽ là người quyết định việc đưa các phương pháp tính toán cũng như kết quả vào bản báo cáo của bạn. Sau đây là ví dụ của mỗi loại kiểm định thống kê và các gợi ý cho bạn khi phiên giải kết quả. Phần cuối của bài này sẽ là một ví dụ về làm thế nào để viết kết quả từ các phân tích thống kê.

4.5. Sử dụng SPSS để kiểm định giả thuyết

4.5.1. So sánh một giá trị trung bình với một giá trị lý thuyết hoặc quần thể

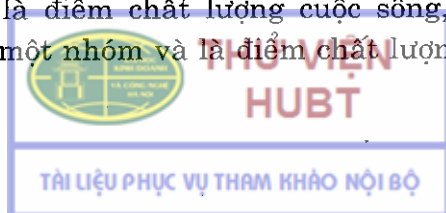
KẾ HOẠCH PHÂN TÍCH – MỘT TRUNG BÌNH

Giả thuyết đầu tiên của chúng ta trong nghiên cứu điều tra tai nạn giao thông quốc gia là:

H_0 : trung bình điểm chất lượng cuộc sống trước khi bị chấn thương của các nạn nhân cũng giống như trong quần thể, điểm là 50.

Kế hoạch phân tích (Analytic plan) được phác thảo như sau:

- i. *Biến phụ thuộc* là điểm chất lượng cuộc sống, không có biến độc lập. Chúng ta chỉ có một nhóm và là điểm chất lượng cuộc sống của toàn bộ,



như vậy câu hỏi là kết quả điểm chất lượng cuộc sống của nghiên cứu này có giống với trung bình của quần thể không? Chúng ta không có số liệu quốc gia chỉ biết là trung bình điểm của toàn bộ quần thể là 50. Như vậy chúng sẽ so sánh mẫu của chúng ta với một giá trị khác hay một giá trị quần thể.

- ii. Điểm chất lượng cuộc sống là *liên tục*.
- iii. Bởi vì liên tục nên đại lượng chọn để tóm tắt số liệu là trung bình và sự biến thiên. Nếu biến có phân bố chuẩn thì sử dụng giá trị trung bình và độ lệch chuẩn nếu không có phân bố chuẩn thì dùng giá trị trung vị và khoảng.
- iv. Các kết quả này không cần thiết phải có bảng.
- v. Sử dụng bảng 1(tr.193), câu hỏi đầu tiên là biến đầu ra (biến điểm chất lượng cuộc sống) có phân bố chuẩn hay không? Do đó trước khi chúng ta chọn loại kiểm định để sử dụng, chúng ta kiểm tra tính chuẩn theo thuật toán ở mục 4.7. Nếu chuẩn thì cột đầu tiên của phần không được bôi đậm trong bảng 1(tr.193) phù hợp, nếu không thì cột đầu tiên của phần bôi đậm sẽ phù hợp. Và kiểm định t cho một giá trị trung bình mẫu nếu như phân bố chuẩn hoặc kiểm định phi tham số tương ứng cho một giá trị trung bình mẫu – kiểm định dấu xếp hạng Wilcoxon – khi phân bố là không chuẩn.
- vi. Các giả định cho kiểm định t cho một giá trị trung bình bao gồm tính chuẩn, xem lại kế hoạch ở trên, và các đơn vị quan sát là độc lập.
- vii. Các đơn vị quan sát trong trường hợp này là người. Chúng ta giả định rằng các đơn vị quan sát là độc lập với nhau từng đôi một (nghĩa là: điểm chất lượng cuộc sống của người này không bị ảnh hưởng bởi điểm chất lượng cuộc sống của người khác). Yêu cầu này là bắt buộc cho tất cả các kiểm định thống kê trong khóa học này. Nếu bạn thấy có hai đơn vị quan sát không độc lập (điểm chất lượng cuộc sống của người này bị ảnh hưởng điểm chất lượng cuộc sống của người khác.) bạn nên tham khảo ý kiến của nhà thống kê.
- viii. Khi điểm chất lượng cuộc sống có phân bố chuẩn, chúng ta sẽ sử dụng kiểm định tham số t cho một giá trị trung bình mẫu.
- ix. Ý nghĩa thống kê được xác định theo quy ước là tại mức $p < 0,05$, sử dụng kiểm định hai phía với giả thuyết không là có sự khác biệt giữa trung bình mẫu và trung bình của quốc gia.

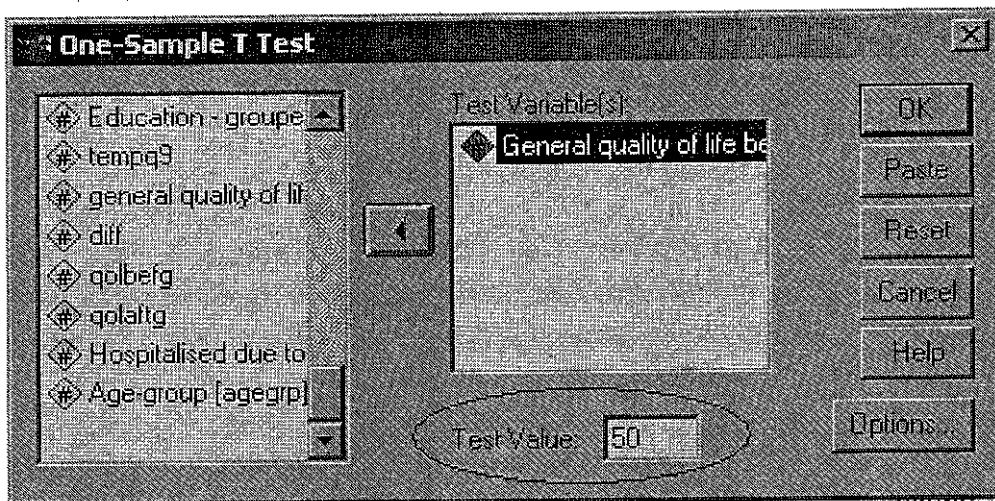
Trong báo cáo bạn nên viết ra một phương pháp giải thích bạn đã chọn kiểm định thống kê như thế nào và tại sao bạn lại chọn như vậy. Với kiểm định này, chúng ta có thể viết là:

Vì điểm chất lượng cuộc sống là biến liên tục với phân bố chuẩn nên kiểm định t cho một giá trị trung bình được sử dụng để kiểm định giả thuyết: H_0 : trung bình điểm chất lượng cuộc sống của các nạn nhân trước khi bị chấn thương cũng giống như quần thể là 50.



SỬ DỤNG SPSS ĐỂ KIỂM ĐỊNH THỐNG KÊ – Kiểm định t một mẫu

1. Từ menu chọn: **Analyse/ Compare Means/One-Sample T Test**. Bạn sẽ có một hộp thoại dạng sau.
2. Từ danh sách các biến đánh dấu vào biến mà bạn muốn phân tích, cụ thể là qol_bef (Quality of Life score before injury– điểm chất lượng cuộc sống trước khi chấn thương), và chuyển nó sang ô Test Variable(s) bằng cách kích vào mũi tên ngang.
3. Viết giá trị mà bạn muốn so sánh với trung bình biến của bạn vào ô Test Value. Trong trường hợp này giá trị so sánh (kiểm định) là điểm trung bình cuộc sống của quần thể người Việt Nam nói chung là 50
4. Kích vào OK.



KẾT QUẢ

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
General quality of life before injury	1692	57.9941	7.68642	.18606

One-Sample Test

	Test Value = 50					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
General quality of life before injury	42.780	1691	.000	7.9941	7.6276	8.3606

PHIÊN GIẢI

Kiểm định thống kê đã được tính. Giá trị trung bình ở đây là 58,0 được so sánh với trung bình quần thể chung là 50. Kiểm định thống kê là sẽ kiểm tra kết quả 58,0 có cao hơn một cách có ý nghĩa thống kê hay không? Kiểm định thống kê, $t = 42,8$ và giá trị p là 0,000 với độ chính xác là ba số 0 sau dấu phẩy (cột *Sig. Column* là giá trị p), dạng ngầm định của spps là kiểm định hai phía. Kết quả này có ý nghĩa, theo quy ước chung, một giá trị p nhỏ hơn hoặc bằng 0,05 được có thể xem là điểm xác định là có ý nghĩa thống kê. Lưu ý rằng, giá trị này được chọn một cách tùy ý. Bậc tự do cũng rất quan trọng, vì giá trị p tương ứng với t của 42,8 sẽ khác nhau phụ thuộc vào bậc tự do.

Hãy cùng quyết định cách để tóm tắt số liệu và phiên giải kết quả phân tích thống kê. Kết quả của bài phân tích phần kiểm định thống kê trong ví dụ trên:

Khi xem xét chất lượng cuộc sống trước khi chấn thương, có một sự khác biệt có ý nghĩa thống kê của trung bình điểm chất lượng cuộc sống của các thành viên tham gia so với điểm trung bình ước tính dựa trên quần thể người Việt Nam chung. ($t_{1691} = 4,8$, $p < 0,001$. Trước khi bị chấn thương, các nạn nhân bị chấn thương giao thông có điểm chất lượng cuộc sống cao hơn 58,0 ($se = 0,2$) so với điểm của quốc gia (50).

Lưu ý: theo quy ước chung bậc tự do của giá trị thống kê t thường được viết theo dạng chỉ số dưới, và không bao giờ viết $p = 0,000$ mặc dù kết quả của máy tính là như vậy. Sẽ chính xác hơn nếu bạn viết kết quả có 3 chữ số sau dấu phẩy và viết $p < 0,001$ kể cả khi $p = 0,000$. Việc phiên giải bằng lời nên đề cập đến việc có ý nghĩa hay không có ý nghĩa thống kê, kiểm định t với bậc tự do và giá

trị p, nếu tìm thấy sự khác biệt, bạn nên đưa vào một câu chỉ ra hướng của sự khác nhau đó. Sẽ tốt hơn nữa khi trình bày số liệu nếu bạn đưa ra giá trị của khoảng tin cậy 95% cho giá trị trung bình quần thể, hoặc ít nhất là sai số chuẩn (S.E). Trung bình mẫu là 58,0 của chúng ta là chỉ là ước lượng của trung bình quần thể vì kết quả này chỉ dựa trên mẫu điều tra của những người đã bị chấn thương giao thông trong khoảng thời gian nghiên cứu mà thôi.

Do cung cấp khoảng tin cậy, ở đây là từ 7,6 đến 8,4, chúng ta có thể thêm rằng: *ước lượng tốt nhất cho sự khác biệt trung bình giữa nạn nhân bị chấn thương giao thông và người Việt Nam chung là 8,0 và chúng ta tin chắc 95% sự khác biệt nằm trong khoảng từ 7,6 tới 8,4.* Kết luận này sẽ cung cấp cho người đọc một vài ý tưởng về tính chính xác của kết quả. Kể cả khi sự khác biệt thực sự chỉ là 7,6 thì đó cũng là sự khác biệt có ý nghĩa. Nếu khoảng tin cậy của chúng ta là từ 0,4 đến 22,7 thì kết quả sẽ kém thuyết phục. Giá trị 8,0 chỉ là ước lượng tốt nhất rút ra từ mẫu nghiên cứu của chúng ta tuy nhiên giá trị thực của quần thể có thể thấp tới 0,4 điểm và sự khác biệt này là không đáng kể, chúng ta nên bàn luận về sự thiếu chính xác của kết quả này trong phần bàn luận của báo cáo.

4.5.2. So sánh trung bình của hai nhóm

KẾ HOẠCH PHÂN TÍCH – HAI TRUNG BÌNH

Xem xét giả thuyết sau đây:

H_0 : điểm trung bình chất lượng cuộc sống trước khi bị chấn thương là như nhau ở hai nhóm nam và nữ.

Kế hoạch phân tích bao gồm những thành phần sau:

Mô tả các biến

Biến phụ thuộc là điểm chất lượng cuộc sống: liên tục.

Biến độc lập là giới tính: phân loại, hai nhóm.

Mô tả mối liên quan

Mối quan hệ sẽ được tóm tắt sử dụng trung bình và sự biến thiên. Trung bình và độ lệch chuẩn nếu biến có phân bố chuẩn, trung vị và khoảng nếu biến không có phân bố chuẩn.

Bảng giá

	Trung bình (mean)	Độ lệch chuẩn (s.d.)
Nam		
Nữ		

Các kiểm định có thể dùng

Sử dụng bảng 1(tr.193) chúng ta có những kiểm định sau:

Kiểm định t không ghép cặp; các giả định là các quan sát độc lập, phương sai đồng nhất và phân bố chuẩn.

Kiểm định Mann–Whitney; các giả định các quan sát độc lập, phương sai đồng nhất.

Chọn kiểm định thống kê

Các giả định được kiểm tra như mục 4.7.

Điểm chất lượng cuộc sống có phân bố chuẩn, phương sai đồng nhất được thỏa mãn.

Tiến hành với kiểm định t không ghép cặp.

Viết báo cáo phương pháp

Phần các phương pháp trong báo cáo của bạn, bạn không cần nêu toàn bộ kế hoạch phân tích, tuy nhiên bạn cần phải chỉ ra loại kiểm định thống kê nào bạn chọn và mối quan hệ nào bạn đã kiểm tra. Phần các phương pháp cho kiểm định thống kê này bạn có thể viết dạng:

Điểm chất lượng cuộc sống trước khi chấn thương có phân bố chuẩn, kiểm định t không ghép cặp hai phía được sử dụng để so sánh điểm trung bình của hai nhóm nam và nữ.

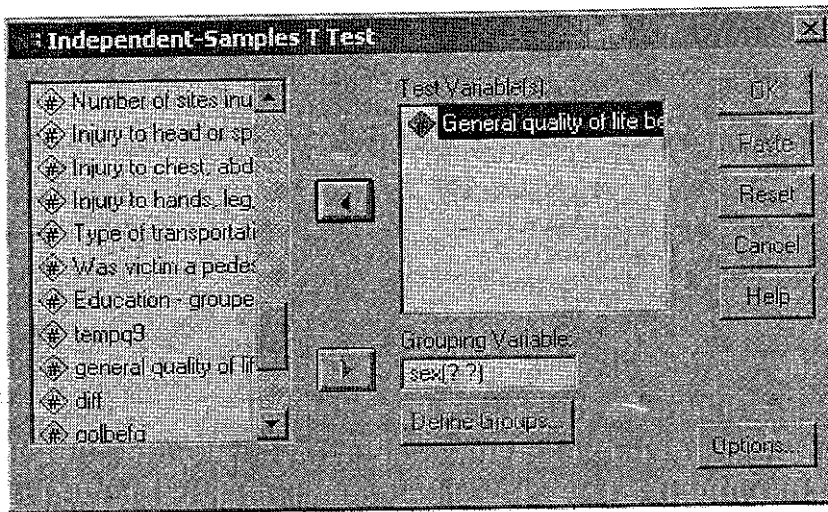


SỬ DỤNG SPSS ĐỂ KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ – So sánh 2 trung bình

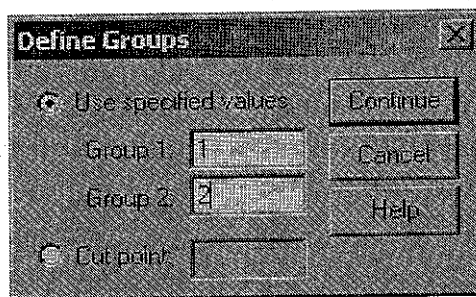
1. Từ menu chọn: **Analyse/Compare Means/Independent–Samples T Test.**

2. Từ danh sách các biến, đánh dấu vào biến phụ thuộc mà bạn muốn phân tích. Trong trường hợp này qol_bef (Quality of Life score before injury), và chuyển nó sang ô Test Variable(s) bằng cách kích vào mũi tên trên.

3. Từ danh sách các biến, đánh dấu vào biến độc lập mà bạn muốn sử dụng có nghĩa là nhóm mà bạn muốn so sánh, trong trường hợp cụ thể này là Sex và chuyển chúng sang ô Grouping Variable bằng cách kích vào mũi tên phía dưới.



4. Bây giờ bạn xác định các loại trong biến độc lập để cho SPSS hiểu được bạn đi so sánh cái gì (có nghĩa là nam và nữ). Để làm điều đó, đánh dấu vào biến độc lập (SEX) và kích vào Define Groups. Bạn sẽ thấy một hộp thoại dạng sau. Bạn cần chỉ ra cho SPSS là mã gì là nam, mã gì là nữ? Để làm được điều này bạn nhập mã cho nam (trong cơ sở dữ liệu được mã là 1) vào trong hộp Group 1 và mã của nữ (trong cơ sở dữ liệu được mã là 2) vào Group 2 sau đó kích Continue.



5. Bây giờ kích OK.

Kết quả sẽ xuất hiện trong một cửa sổ kết quả như sau:

KẾT QUẢ

Group Statistics



	sex	N	Mean	Std. Deviation	Std. Error Mean
General quality of life before injury	male	1089	58.0909	7.79659	.23626
	female	603	57.8192	7.48657	.30488

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference		
										Lower	Upper
General quality of life before injury	Equal variances assumed	1.684	.195	.696	1690	.486	.2717	.39023	-4.9371	1.03705	
	Equal variances not assumed			.704	1285.805	.481	.2717	.36571	-4.8501	1.02835	

PHIÊN GIẢI

Có 1089 nam và 603 nữ trong mẫu chấn thương giao thông của chúng ta. Điểm trung bình của nam và nữ khá giống nhau 58,1 và 57,8; sự khác biệt về điểm chất lượng cuộc sống giữa nam và nữ là 0,3. Sự khác biệt này là không đủ lớn để chúng ta cho rằng có sự khác biệt về trung bình điểm của nam và nữ. Liệu kết quả của kiểm định thống kê có khẳng định điều này không? Lưu ý rằng kết quả của phần mềm cung cấp cho bạn các phép tính toán cho hai tình huống: thứ nhất là giả sử *phương sai bằng nhau*, và thứ hai là *phương sai không bằng nhau*. Bảng 1(tr.193) cho chúng ta biết rằng chúng ta nên kiểm tra các giả định về sự đồng nhất của phương sai của giữa các nhóm, điều này sẽ bàn luận trong mục 4.7. Tuy nhiên kiểm định t không ghép cặp cũng vẫn được thực hiện và cho chúng ta kết quả trên đây.

Kiểm định Levene ở phần đầu tiên của bảng cần phải được phiên giải riêng biệt với các phần còn lại của bảng bắt đầu bắt đầu từ cột t. Kiểm định Levene cho kết quả không có ý nghĩa ($p = 0,195$), do đó chúng ta giả định rằng phương sai xấp xỉ bằng nhau, chúng ta sử dụng kết quả kiểm định t với “giả định phương sai bằng nhau”. Nếu kiểm định Levene có ý nghĩa ($p < 0,05$), thì bậc tự do cần hiệu chỉnh khi phiên giải cho kiểm định t, sự hiệu chỉnh này diễn ra đối với bậc tự do và nó kéo theo giá trị p cũng thay đổi.

Gộp toàn bộ các kết quả phía trên, chúng ta có thể đưa vào trong báo cáo các thông tin sau:

Khi xem xét điểm chất lượng cuộc sống trước khi chấn thương, chúng ta không có đủ bằng chứng về sự khác biệt của trung bình chất lượng cuộc sống giữa nam và nữ ($t_{1690} = 0,5$, $p = 0,486$). Trung bình sự khác biệt về điểm của nam và nữ là 0,3 điểm với khoảng tin cậy 95% là (-0,5 đến +1,0).

Bậc tự do của kiểm định này là 1690 (kiểm định t không ghép cặp) so với bậc tự do của kiểm định trước là 1691 (kiểm định t cho một mẫu) bởi vì trong trường hợp này chúng ta ước lượng cho hai giá trị trung bình mẫu còn trong trường hợp trước là ước lượng cho một giá trị trung bình.

4.5.3. So sánh giá trị trung bình nhiều hơn hai nhóm

LẬP KẾ HOẠCH PHÂN TÍCH – NHIỀU HƠN HAI NHÓM

Xét giả thuyết sau:

H_0 : trung bình điểm chất lượng cuộc sống trước khi bị chấn thương giống nhau ở tất cả các mức trình độ học vấn.

Giả thuyết này tương tự như như giả thuyết hai nhóm ở trên nhưng bây giờ là so sánh nhiều hơn hai nhóm. Bạn nên lập kế hoạch phân tích như sau:

Mô tả các biến

Một biến phụ thuộc là điểm chất lượng cuộc sống, liên tục.

Một biến độc lập là trình độ học vấn: phân loại, 4 nhóm.

Mô tả mối liên quan

Mối quan hệ sẽ được tóm tắt là trung bình đại số và sự phân tán: trung bình (s.d) nếu biến có phân bố chuẩn, trung vị (khoảng) nếu biến không có phân bố chuẩn.

Bảng giả

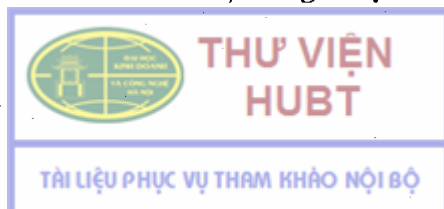
Trình độ học vấn	Điểm chất lượng cuộc sống	
	Trung bình (mean)	Độ lệch chuẩn (sd)
Còn nhỏ		
Cấp 1		
Cấp 2		
Cấp 3 trở lên		

Xác định các kiểm định thống kê có thể dùng

Sử dụng các kiểm định thống kê trong bảng 1(tr.193).

Kiểm định ANOVA một chiều, các giả định là các quan sát độc lập, phân bố chuẩn và phương sai đồng nhất.

Kiểm định ANOVA Kruskal-Wallis; các giả định là các quan sát độc lập và phương sai đồng nhất.



Chọn kiểm định thống kê

Các giả định được kiểm tra theo mục 4.7.

Điểm chất lượng cuộc sống có phân bố chuẩn và thỏa mãn phương sai đồng nhất.

Sử dụng kiểm định ANOVA một chiều.

Viết báo cáo phương pháp

Phần các phương pháp của bạn trong kiểm định nên viết báo cáo có dạng:

Điểm chất lượng cuộc sống trước khi bị chấn thương được chứng minh là có phân bố chuẩn nên chúng ta có thể sử dụng phân tích phương sai ANOVA hai phía để so sánh trung bình điểm chất lượng cuộc sống giữa các mức trình độ học vấn (4 nhóm).

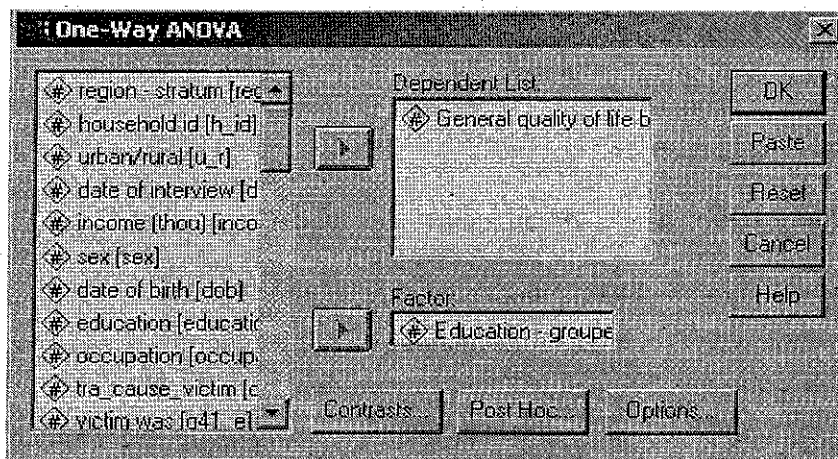


SỬ DỤNG SPSS ĐỂ KIỂM ĐỊNH GIẢ THUYẾT – So sánh nhiều trung bình

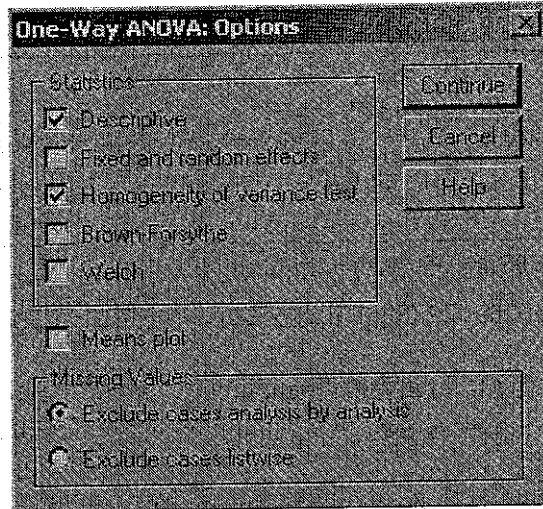
1. Từ menu chọn: **Analyse/Compare Means/One-Way ANOVA**. Bạn sẽ có hộp thoại có dạng sau.

2. Từ danh sách các biến đánh dấu vào biến phụ thuộc mà bạn muốn phân tích, trong trường hợp này là qol_bef (Quality of Life score before injury– điểm chất lượng cuộc sống trước khi bị chấn thương), và chuyển nó vào ô Dependent List bằng cách kích vào mũi tên phía trên.

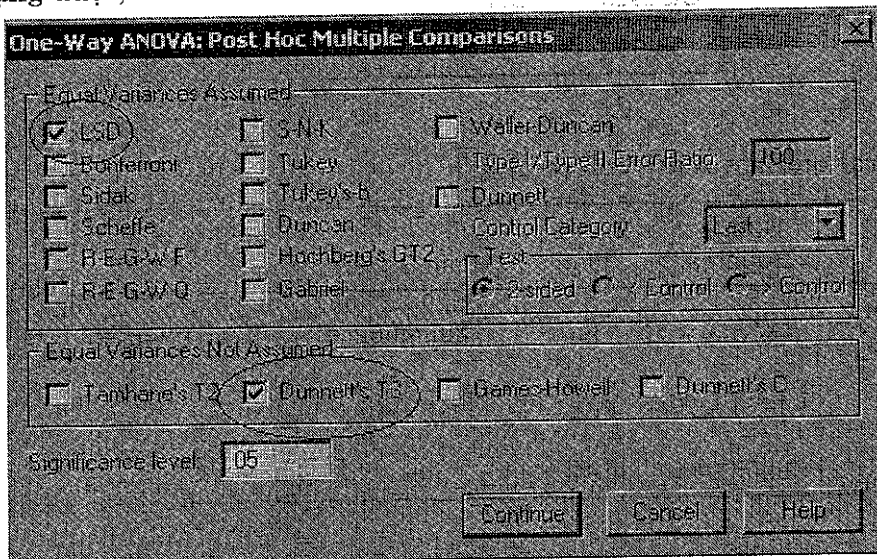
3. Từ danh sách các biến, đánh dấu vào biến độc lập mà bạn muốn sử dụng (có nghĩa là các nhóm mà bạn muốn so sánh). Trong trường hợp này là education, và chuyển nó sang ô Factor bằng cách kích vào mũi tên phía dưới.



4. Để có kết quả về thống kê mô tả (điểm trung bình của các nhóm) và cho mỗi kiểm định thống kê phương sai đồng nhất hãy kích vào Options. Bạn có hộp thoại dạng như sau, chú ý cần chọn mục Descriptive và Homogeneity of variance test. Sau đó Continue.



5. Để so sánh từng cặp bạn cần kích vào Post Hoc, bạn sẽ có một hộp thoại dạng sau; bạn có thể chọn bất cứ phương pháp so sánh cặp nào. Tuy nhiên chúng ta nên dùng LSD nếu bạn có các phương sai bằng nhau và Dunnett's T3 nếu các phương sai của bạn không bằng nhau, đó là các kiểm định "an toàn" có thể sử dụng được, sau đó kích vào Continue.



6. Bây giờ kích vào OK.

Kết quả đầu ra của bạn sẽ xuất hiện ở một cửa sổ riêng biệt. Cửa sổ đó có dạng như sau:



KẾT QUẢ

Oneway

Descriptives

General quality of life before injury

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Children	83	63.8554	5.78925	.63545	62.5913	65.1195	50.00	76.00
Less than secondary	469	55.3369	7.04667	.32538	54.6975	55.9763	34.00	75.00
Secondary	987	58.1651	7.62869	.24282	57.6886	58.6417	36.00	85.00
More than Secondary	153	61.8562	7.28320	.58881	60.6929	63.0195	45.00	80.00
Total	1692	57.9941	7.68642	.18686	57.6276	58.3606	34.00	85.00

Test of Homogeneity of Variances

General quality of life before injury

Levene Statistic	Df1	df2	Sig.
3.900	3	1688	.009

ANOVA

General quality of life before injury

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	8473.986	3	2824.662	52.148	.000
Within Groups	91431.955	1688	54.166		
Total	99905.941	1691			

Post Hoc Tests

Multiple Comparisons

Dependent Variable: General quality of life before injury

	(I) Education - grouped	(J) Education - grouped	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	Children	Less than secondary	8.5185(*)	.87841	.000	6.7398	10.2375
		Secondary	5.6903(*)	.84112	.000	4.0405	7.3400
		More than secondary	1.9992(*)	1.00331	.046	.0314	3.9671
	Less than secondary	Children	-8.5185(*)	.87841	.000	-10.2375	-6.7996
		Secondary	-2.8283(*)	.41276	.000	-3.6378	-2.0187
		More than secondary	-6.5193(*)	.68521	.000	-7.8633	-5.1754
	Secondary	Children	-5.6903(*)	.84112	.000	-7.3400	-4.0405
		Less than secondary	2.8283(*)	.41276	.000	2.0187	3.6378
		More than secondary	-3.6911(*)	.63946	.000	-4.9463	-2.4369
	More than secondary	Children	-1.9992(*)	1.00331	.046	-3.9671	-.0314
		Less than secondary	6.5193(*)	.68521	.000	5.1754	7.8633
		Secondary	3.6911(*)	.63946	.000	2.4369	4.9463
Duncan T3	Children	Less than secondary	8.5185(*)	.71392	.000	6.6123	10.4248
		Secondary	5.6903(*)	.68027	.000	3.8685	7.5121
		More than secondary	1.9992	.86831	.124	-.3016	4.3000
	Less than secondary	Children	-8.5185(*)	.71392	.000	-10.4248	-6.6123
		Secondary	-2.8283(*)	.40600	.000	-3.8685	-1.7590
		More than secondary	-6.5193(*)	.87274	.000	-8.3027	-4.7359
	Secondary	Children	-5.6903(*)	.68027	.000	-7.5121	-3.8685
		Less than secondary	2.8283(*)	.40600	.000	1.7590	3.8685
		More than secondary	-3.6911(*)	.63692	.000	-5.3923	-1.9990
	More than secondary	Children	-1.9992	.86831	.124	-4.3000	-.3016
		Less than secondary	6.5193(*)	.67274	.000	4.7359	8.3027
		Secondary	3.6911(*)	.63692	.000	1.9990	5.3923

*. The mean difference is significant at the .05 level.

PHIÊN GIẢI

Nhìn vào kết quả ta thấy điểm có xu hướng tăng lên cùng với sự tăng lên của trình độ học vấn (trung bình từ 55 đến 64). Kiểm định phân tích phương sai thực chất là có hai giai đoạn. Bởi vì ở đây so sánh nhiều hơn hai nhóm nên giai đoạn đầu là kiểm định toàn bộ, xét giả thuyết sau đây:

H_0 : Tất cả bốn trung bình đều bằng nhau.

H_1 : Có ít nhất một trung bình khác với các trung bình còn lại.

Kiểm định ANOVA có thể tóm tắt nhau sau:

Có đủ bằng chứng để nói rằng có sự khác biệt điểm chất lượng cuộc sống giữa các trình độ học vấn ($F_{3,1688} = 52,1, p < 0,001$).

Vì giai đoạn 1 của kiểm định có ý nghĩa thống kê nên chúng ta có thể tiến hành so sánh từng cặp trung bình. Một trong các phương pháp đó là kiểm định t mà bạn đã được học ở trên, ngoài ra chúng ta có nhiều phương pháp khác để kiểm định từng cặp trung bình. Kiểm định LSD, dựa trên cơ bản của kiểm định t ghép cặp và kiểm định Duncan vẫn hay được sử dụng trong các nghiên cứu.



Tuy nhiên, chúng yêu cầu giả thiết phương sai đồng nhất phải thoả mãn (xem mục 4.7). Nếu giả định đó không thoả mãn thì dùng kiểm định Dunnett's T3.

Bởi vì giai đoạn đầu tiên của chúng ta có ý nghĩa thống kê nên chúng ta kỳ vọng có ít nhất một cặp so sánh có ý nghĩa. Trong trường hợp này đó là kết quả đầu ra ở trên. Các giả định phương sai đồng nhất không thoả mãn ($p = 0,0009$) nên chúng ta nhìn vào kết quả kiểm định Dunnett's T3 post-hoc. Và báo cáo kết quả của bạn có thể là như sau:

Các kết quả chỉ ra rằng tất cả các cặp so sánh đều có ý nghĩa thống kê từng đôi một ($p < 0,05$). Điểm trung bình cuộc sống ở tuổi còn nhỏ có trung bình cao hơn cấp 1 và cấp 2 (Kiểm định Dunnett's T3 với $p < 0,001$). Chất lượng cuộc sống giảm ở các mức trình độ học vấn thấp.

Một tình huống khác

Kiểm định ANOVA có thể được sử dụng để kiểm định sự khác biệt của nhiều nhóm. Sau đây là đầu ra từ một kiểm định ANOVA cho giả thuyết sau:

H_0 : Trung bình đại số của điểm chất lượng cuộc sống trước khi bị chấn thương là giống nhau ở tất cả các vùng nghiên cứu.

Oneway

Descriptives

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
ĐL	199	58,5678	8,22562	,56310	57,4180	59,7177	34,00	78,00
ĐVL	55	57,1636	8,19160	1,10456	54,9491	59,3781	43,00	74,00
Đ	194	57,9485	7,35181	,52783	56,9074	58,9895	39,00	73,00
ĐL	254	58,8976	7,36012	,46181	57,9881	59,8071	38,00	85,00
cc	151	58,2914	8,62986	,70229	56,9037	59,6790	36,00	78,00
gh	154	57,7208	7,69050	,61972	56,4965	58,9451	36,00	76,00
se	289	58,2363	7,43350	,43728	57,3747	58,0959	35,00	80,00
ĐL	396	57,0808	7,46135	,37444	56,3447	57,8170	36,00	82,00
Total	1692	57,9941	7,68642	,18686	57,5276	58,3606	34,00	85,00

Test of Homogeneity of Variances

General quality of life before injury			
Levene Statistic	df1	df2	Sig.
1,746	7	1684	,094

ANOVA

General quality of life before injury					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	683,170	7	97,596	1,656	,116
Within Groups	9922,771	1684	58,921		
Total	99905,941	1691			



Từ phần mô tả chúng ta thấy rằng điểm chất lượng cuộc sống ở các cùng có sự khác nhau rất ít. Kết quả của kiểm định ANOVA chỉ ra rằng sự khác biệt không có ý nghĩa thống kê. Như vậy, phiên giải kết quả chỉ ra rằng không có đủ bằng chứng thống kê về một mối liên quan nhưng bạn nên cung cấp các kết quả và bình luận cho phép người đọc quyết định liệu có nên chỉ kết luận là có ít nhất một trung bình khác các nhóm còn lại khác. Mặc dù SPSS có khả năng tiến hành cho bạn các kiểm định so sánh từng cặp như trong ví dụ trên, nhưng vì kết quả của kiểm định ANOVA không có ý nghĩa thống kê nên trong phần phiên giải kết quả, bạn KHÔNG nên nói về giai đoạn hai và bỏ qua cả kết quả của kiểm định LSD và Dunnett's T3. Nếu thống kê F cho kết quả có ý nghĩa thống kê thì kiểm định sự đồng nhất của phương sai chỉ ra rằng chúng ta sẽ sử dụng kết quả của kiểm định LSD để kết luận cho việc so sánh từng cặp.

Kết luận trong trường hợp này:

Không đủ bằng chứng để nói rằng trung bình điểm chất lượng cuộc sống có sự khác biệt giữa các cùng sinh thái trong nghiên cứu này ($F_{7,1684} = 1,7, p = 0,116$).

4.5.4. So sánh đo lường lặp lại trên cùng một đơn vị – so sánh trung bình

LẬP KẾ HOẠCH PHÂN TÍCH – CÁC TRUNG BÌNH LẶP LẠI

Xét giả thuyết sau:

H_0 : trung bình điểm chất lượng cuộc sống sau khi chấn thương giao thông không thấp hơn trước khi chấn thương giao thông.

Bạn nên thực hiện các bước sau khi lập kế hoạch phân tích.

Mô tả các biến

Một biến phụ thuộc là điểm chất lượng cuộc sống, liên tục, lặp lại theo thời gian.

Mô tả mối liên quan

Mối liên quan có thể được tóm tắt như là các trung bình và sự phân tán: trung bình (s.d) nếu biến có phân bố chuẩn, trung vị (khoảng) nếu biến không có phân bố chuẩn.

Bảng giá

	Trung bình (mean)	Sự biến thiên(s.d.)
QoL trước khi chấn thương		
QoL sau khi chấn thương		



Xác định các loại kiểm định có thể dùng

Sử dụng các kiểm định trong bảng 1(tr.193).

Kiểm định t ghép cặp, giả định các quan sát độc lập, sự khác nhau có tính chuẩn.

Kiểm định dạng dấu Wilcoxon; các giả định: các quan sát độc lập.

Chọn lựa kiểm định thống kê cuối cùng.

Các giả định được kiểm tra theo mục 4.7.

Sự khác nhau về điểm chất lượng cuộc sống trước và sau chấn thương có phân bố chuẩn, thực hiện kiểm định t ghép cặp.

Giả thuyết mong muốn tìm ra là điểm chất lượng cuộc sống có thấp hơn sau khi bị chấn thương không nên sử dụng kiểm định một phía.

Viết báo cáo phương pháp

Phần các phương pháp của bạn trong kiểm định này sẽ được viết dưới dạng

Sự khác biệt về điểm chất lượng cuộc sống trước và sau khi bị chấn thương đã được kiểm tra là có phân bố chuẩn nên kiểm định t ghép cặp là phù hợp để so sánh trung bình điểm qua thời gian. Do mong muốn điểm chất lượng cuộc sống trước khi bị chấn thương sẽ cao hơn so với sau khi chấn thương nên chúng ta sử dụng kiểm định một phía.

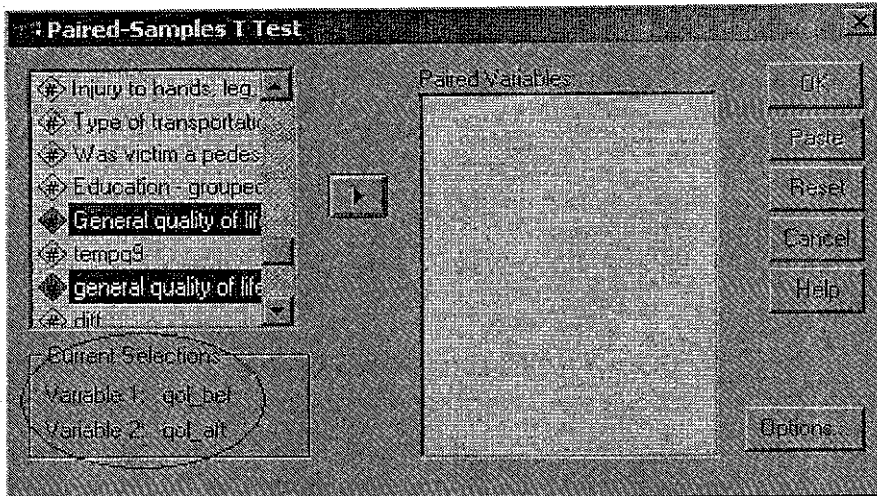


SỬ DỤNG SPSS ĐỂ KIỂM ĐỊNH GIẢ THUYẾT – Kiểm định t ghép cặp

1. Để thực hiện kiểm định t ghép cặp từ thanh menu chọn: **Analyse/Compare Means/Paired-Samples T Test**. Bạn sẽ có một hộp thoại như dưới đây

2. Từ danh sách các biến, đánh dấu vào cặp biến thể hiện độ đo hiện nay và độ đo lặp lại, trong trường hợp này là qol_bef (Quality of Life score before injury– điểm chất lượng cuộc sống trước khi chấn thương) và qol_aft (Quality of Life score after injury– điểm chất lượng cuộc sống sau khi bị chấn thương). Bạn sẽ thấy chúng xuất hiện như là biến 1 và biến 2 trong ô Current Selections.

3. Bây giờ hãy chuyển cặp biến vào ô Paired Variables bằng cách kích vào mũi tên. Bạn sẽ thấy chúng chuyển sang ô Paired Variables được liên kết với nhau bằng dấu chấm gạch.



4. Kịch OK.

KẾT QUẢ

T-Test

Paired Samples Statistics

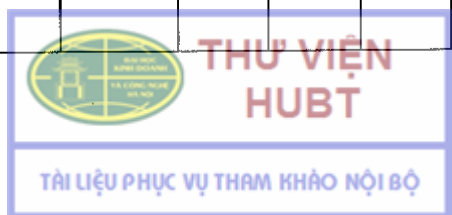
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	General quality of life before injury	60.4054	1692	7.67448	.18657
	general quality of life after injury	54.6820	1692	9.98208	.24957

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	General quality of life before injury & general quality of life after injury	1692	.787	.000

Paired Samples Test

		Paired Differences				t	df	Sig. (2-tailed)	
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower				Upper
Pair 1	General quality of life before injury - general quality of life after injury	5.7234	6.16665	.14992	5.4294	6.0174	38.177	.000	



PHIÊN GIẢI

Đầu ra của kiểm định t được phiên giải theo hướng của kiểm định t không ghép cặp (xem phần 4.5.2). Bạn có thể viết các kết quả của bạn có dạng như sau:

Trung bình điểm giảm 5,7 đơn vị điểm theo thời gian (từ 60,4 xuống 54,7) với khoảng tin cậy 95% là 5,4 đến 6,0. Kết quả này có ý nghĩa thống kê ($t_{1691} = 38,2, p < 0,001$).

Lưu ý rằng SPSS không cung cấp cho bạn các kết quả của kiểm định một phía. Bạn sẽ phải phiên giải kiểm định thống kê t một cách phù hợp bằng cách sử dụng các bảng thống kê để xác định mức ý nghĩa. Tham khảo Thống kê Y tế Công cộng phần Thống kê cơ bản để biết thêm chi tiết về kiểm định này. Ở đây, giá trị p cho kiểm định một và hai phía là tương tự nhau vì giá trị $p < 0,001$. Tuy nhiên, một điều rất quan trọng là cần phải hiệu chỉnh giá trị p mà SPSS cung cấp nếu bạn muốn kết luận cho kiểm định một phía.

4.5.5. So sánh đo lường lặp lại trên cùng một đơn vị – so sánh trung vị

LẬP KẾ HOẠCH PHÂN TÍCH – TRUNG BÌNH LẶP LẠI

Xét giả thuyết như trên:

H_0 : trung bình điểm chất lượng cuộc sống của các nạn nhân trước và sau khi bị chấn thương giao thông không khác nhau.

Các bước trong lập kế hoạch phân tích:

Các kiểm định thống kê có thể dùng là hai, một kiểm định tham số (kiểm định t ghép cặp) và một kiểm định phi tham số (kiểm định dấu xếp hạng Wilcoxon). Nếu sự khác biệt về điểm chất lượng cuộc sống có phân bố chuẩn thì dùng kiểm định t ghép cặp. Tuy nhiên, kết quả cũng có giá trị tương đương nếu bạn dùng kiểm định dấu hạng Wilcoxon. Và nếu sự khác biệt KHÔNG có phân bố chuẩn thì chúng ta phải dùng kiểm định dấu hạng Wilcoxon.

Viết báo cáo phương pháp

Nếu kiểm định dấu xếp hạng Wilcoxon là kiểm định phù hợp thì báo cáo phương pháp sẽ có dạng:

Chúng ta sử dụng kiểm định Wilcoxon để kiểm định sự thay đổi điểm chất lượng cuộc sống qua thời gian do sự khác biệt về điểm chất lượng cuộc sống trước và sau khi chấn thương không có phân bố chuẩn.

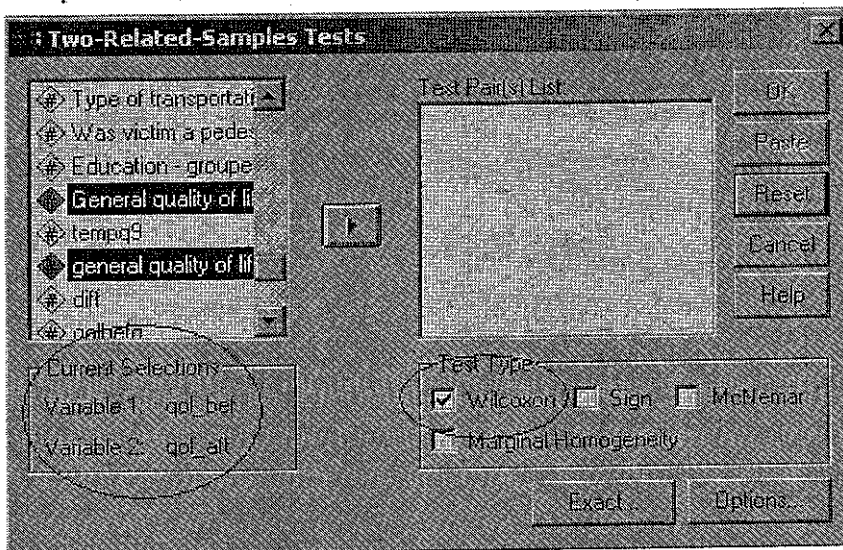


SỬ DỤNG SPSS ĐỂ KIỂM ĐỊNH THỐNG KÊ – Kiểm định dấu xếp hạng

Để sử dụng SPSS tiến hành kiểm định sự khác biệt điểm chất lượng cuộc sống trước và sau khi bị chấn thương bạn cần phải tạo ra một biến mới. Biến mới được tính từ điểm khác biệt trước và sau chấn thương cho mỗi người. Để làm điều này, thực hiện theo các bước trong bài 2 phần 2.3.2.1 về thao tác với dữ liệu. Tiếp theo, bạn thực hiện các bước như bài 3 phần 3.5.3 thống kê mô tả để tính trung vị sự khác biệt nếu như SPSS không thực hiện điều đó cho bạn các việc trên như một phần trong kết quả đầu ra của kiểm định xếp hạng dấu Wilcoxon.

1. Từ thanh thực đơn, chạy kiểm định xếp hạng dấu: **Analyse/Nonparametric Tests/2 Related Samples**. Một hộp thoại sẽ được mở ra.

2. Từ danh sách các biến, đánh dấu vào cặp biến thể hiện các đo lường lặp lại, trong trường hợp này là qol_bef (Quality of Life score before injury–điểm chất lượng cuộc sống trước khi bị chấn thương) và qol_aft (Quality of Life score after injury–điểm chất lượng cuộc sống sau khi bị chấn thương). Bạn sẽ thấy chúng xuất hiện như là biến 1 và biến 2 trong ô Current Selections.



3. Bây giờ bạn chuyển các biến sang ô Paired Variables bằng cách kích vào mũi tên. Bạn sẽ thấy chúng chuyển qua ô Paired Variables và được liên kết qua dấu chấm gạch.

4. Hãy chắc chắn rằng lựa chọn Wilcoxon được chọn trong hộp Test Type.

5. Kích OK.



KẾT QUẢ

Summarize

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
DIFF	1692	98.3%	29	1.7%	1721	100.0%

Report

DIFF

N	Median	Minimum	Maximum
1692	-6.0000	-26.00	13.00

NPar Tests

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
general quality of life after injury - General quality of life before injury	Negative Ranks	1330(a)	853.38	1134902.50
	Positive Ranks	240(b)	409.34	98242.50
	Ties	122(c)		
	Total	1692		

- a. general quality of life after injury < General quality of life before injury
 b. general quality of life after injury > General quality of life before injury
 c. general quality of life after injury = General quality of life before injury

Test Statistics(b)

	general quality of life after injury - General quality of life before injury
Z	-28.870(a)
Asymp. Sig. (2-tailed)	.000

- a. Based on positive ranks.
 b. Wilcoxon Signed Ranks Test

PHIÊN GIẢI

Kết quả của kiểm định dấu xếp hạng Wilcoxon được phiên giải giống như kiểm định Mann-Whitney U (xem sau). Gộp tất cả các kết quả chúng ta sẽ viết báo cáo như sau:

Chúng ta có đủ bằng chứng để kết luận rằng điểm chất lượng cuộc sống qua thời gian có sự khác biệt có ý nghĩa thống kê ($Z = -28.9$, $n = 1692$, $p < 0,001$). Trung vị của sự khác biệt của điểm chất lượng cuộc sống sau khi bị chấn thương thấp hơn 6 điểm so với điểm chất lượng cuộc sống trước khi bị chấn thương (khoảng từ -26 đến $+13$).

4.5.6. So sánh đo lường lặp lại trên cùng một đơn vị – các tỷ lệ**LẬP KẾ HOẠCH PHÂN TÍCH – CÁC TỶ LỆ LẶP LẠI**

Các nhà nghiên cứu cũng xem xét giả thuyết sau:

H_0 : điểm chất lượng cuộc sống thấp (điểm được phân loại) sau khi chấn thương bằng với điểm chất lượng cuộc sống thấp trước khi bị chấn thương.

Bạn nên theo các bước lập kế hoạch phân tích số liệu như sau:

Mô tả các biến

Một biến phụ thuộc là điểm chất lượng cuộc sống (được phân loại)–phân loại nhị thức; lặp lại qua thời gian.

Mô tả mối liên quan

Mối tương quan được tóm tắt như số lượng và các tỷ lệ.

Bảng giả

	Số lượng	Tỷ lệ
QoL trước khi bị chấn thương		
QoL sau khi chấn thương		

Xác định các kiểm định thống kê có thể dùng

Sử dụng bảng 2(tr.194) khi biến phụ thuộc là định danh, kiểm định thống kê là khi bình phương McNemar.

Quyết định kiểm định thống kê cuối cùng

Các giả định được kiểm tra như tại mục 4.7.

Các giả định không thỏa mãn.

Thực hiện kiểm định khi bình phương McNemar.

Viết báo cáo phương pháp

Phần các phương pháp của bạn cho kiểm định thống kê này, bạn nên viết báo cáo có dạng như sau:

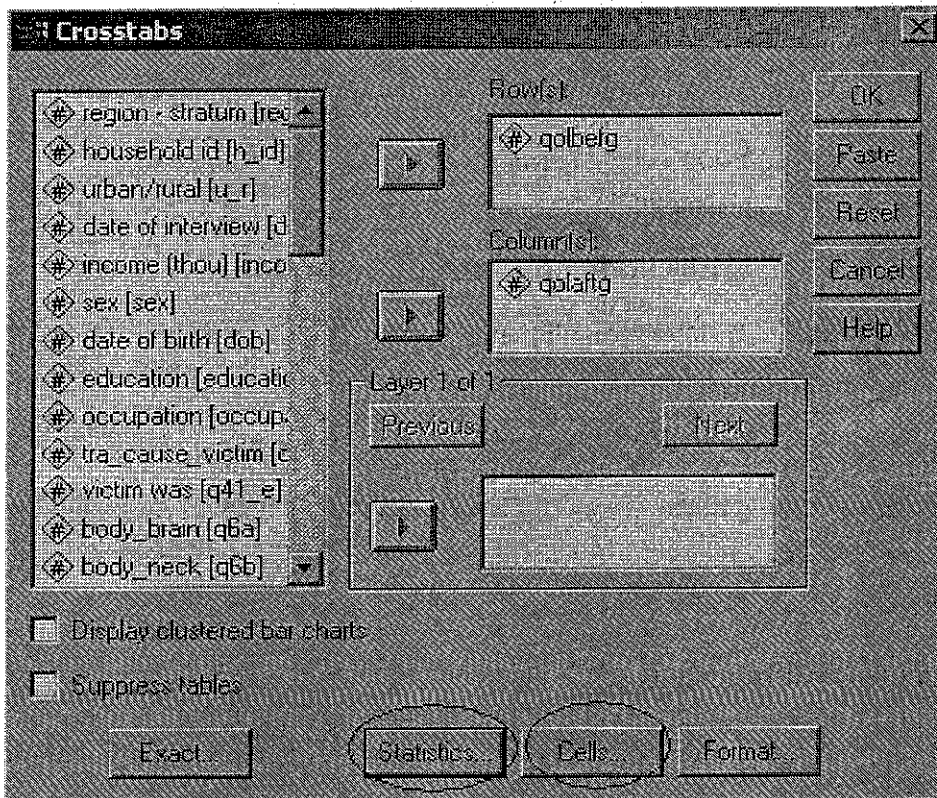
Để kiểm tra sự khác nhau về tỷ lệ người có điểm chất lượng cuộc sống thấp (điểm <50) của trước và sau khi chấn thương chúng ta sử dụng kiểm định khi bình phương McNemar. Nếu chúng ta muốn kiểm tra liệu điểm chất lượng cuộc sống trước khi chấn thương sẽ cao hơn sau khi bị chấn thương thì chúng ta sử dụng kiểm định một phía.



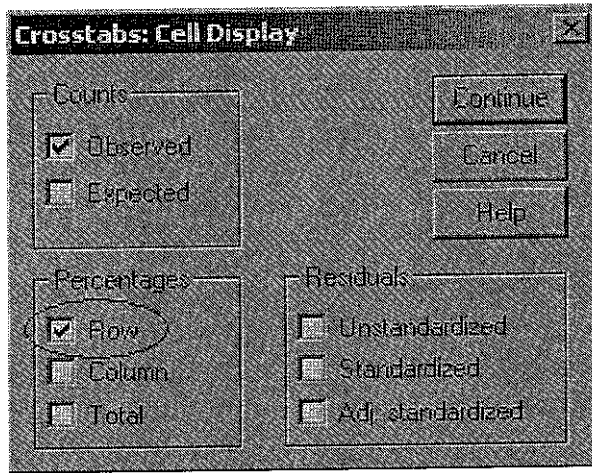


DÙNG SPSS ĐỂ KIỂM ĐỊNH THỐNG KÊ – Kiểm định Khi bình phương McNEMAR

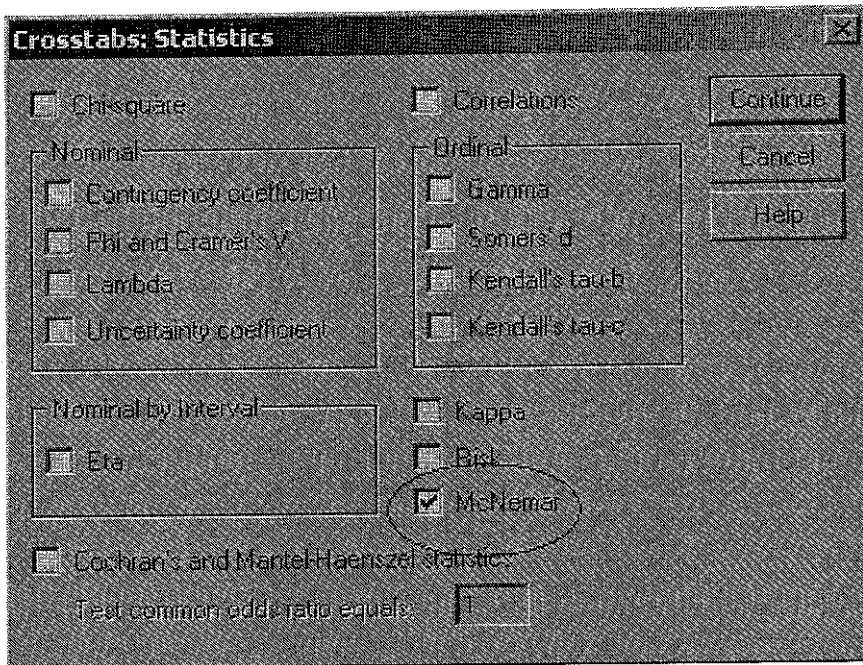
1. Từ menu chọn: **Analyse/Descriptive Statistics/Crosstabs**. Bạn sẽ có một hộp thoại như sau.
2. Từ danh sách các biến, đánh dấu vào biến tương ứng với đo lường lần một, trong trường này là qolbefg (Quality of Life before injury – điểm chất lượng cuộc sống trước khi chấn thương – (được nhóm)), và chuyển nó sang ô Row(s) bằng cách kích vào dấu mũi tên.
3. Từ danh sách các biến, đánh dấu vào biến độc lập mà bạn muốn phân tích trong trường hợp này là qolaftg (Quality of Life after injury điểm chất lượng cuộc sống sau khi chấn thương (đã được nhóm), và chuyển nó sang ô Column(s) bằng cách kích vào mũi tên.



4. Để hiển thị tỷ lệ những người có các loại điểm chất lượng cuộc sống sau khi chấn thương trên điểm chất lượng cuộc sống trước khi bị chấn thương hay bạn muốn hiển thị tỷ lệ HÀNG, sau đó kích Cells và có một ô mới xuất hiện có dạng như sau: đánh dấu vào hộp thoại. Chọn Row trong hộp thoại sau đó kích Continue.



5. Để yêu cầu SPSS tính kiểm định thống kê Khi bình phương McNemar, bạn kích Statistics và một hộp thoại mới dạng sau sẽ xuất hiện. Chọn McNemar và sau đó kích Continue.



6. Bây giờ kích OK

KẾT QUẢ



Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
QOLBEFG * QOLAFTG	1692	98.3%	29	1.7%	1721	100.0%

QOLBEFG * QOLAFTG Crosstabulation

			QOLAFTG		Total
			Suboptimal QOL	Adequate QOL	
QOLBEFG	Suboptimal QOL	Count	145	13	158
		% within QOLBEFG	91.8%	8.2%	100.0%
	Adequate QOL	Count	420	1108	1528
		% within QOLBEFG	27.8%	72.2%	100.0%
Total		Count	571	1121	1692
		% within QOLBEFG	33.7%	66.3%	100.0%

Chi-Square Tests

	Value	Exact Sig. (2-sided)
McNemar Test		.000(a)
N of Valid Cases	1692	

a. Binomial distribution used.

PHIÊN GIẢI

Trong những người có điểm chất lượng cuộc sống đạt chuẩn trước khi chấn thương thì có 28% có điểm chất lượng cuộc sống xuống mức thấp sau khi bị chấn thương. Có 8% những người có điểm chất lượng cuộc sống thấp trước khi chấn thương đã nâng lên mức đạt chuẩn sau khi bị chấn thương. Kiểm định thống kê cho sự thay đổi của các độ đo của một biến phân loại là kiểm định liệu có sự thay đổi khác nhau qua thời gian hay không? Có nghĩa là có phải là người có điểm chất lượng cuộc sống thấp hơn sau khi bị chấn thương nhiều hơn những người có điểm chất lượng cuộc sống cao hơn sau khi chấn thương? Một kiểm định thống kê tương tự có thể được áp dụng cho trường hợp biến đầu ra có nhiều loại.

Về mặt thống kê, phân bố của sự tăng lên hay giảm đi của điểm chất lượng cuộc sống là khác nhau nên báo cáo của bạn có thể viết dưới dạng:

28% nạn nhân có điểm chất lượng cuộc sống đầy đủ trước khi chấn thương đã giảm đi xuống mức thấp sau khi bị chấn thương. Tỷ lệ giảm sút điểm chất lượng cuộc sống cao hơn một cách có ý nghĩa thống kê so với tỷ lệ tăng điểm chất lượng cuộc sống (kiểm định McNemar với 2 bậc tự do, $p < 0,001$).



4.5.7. So sánh trung vị của hai nhóm

LẬP KẾ HOẠCH PHÂN TÍCH – HAI TRUNG VỊ

Xét giả thuyết tiếp:

H_0 : trong những người nằm viện thì trung bình ngày nằm viện của nhóm nạn nhân đi bộ và dùng các loại xe là như nhau.

Kế hoạch phân tích bao gồm các bước sau:

Mô tả các biến

Biến phụ thuộc là số ngày nằm viện, biến liên tục.

Biến độc lập là đi bộ, biến phân loại và có hai nhóm.

Mô tả mối liên quan

Mối liên quan có thể được tóm tắt thông qua trung bình và sự biến thiên: trung bình và độ lệch chuẩn nếu biến có phân bố chuẩn; trung vị và khoảng nếu biến không có phân bố chuẩn.

Bảng giả

	Điểm chất lượng cuộc sống	
	Trung bình (mean)	Sự biến thiên (s.d.)
Đi bộ		
Không đi bộ		

Xác định kiểm định thống kê có thể dùng

Sử dụng các kiểm định thống kê trong bảng 1(tr.193) là:

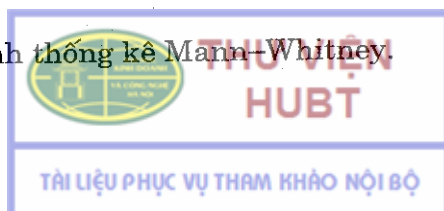
- Kiểm định t không ghép cặp, các giả định là các quan sát độc lập, phương sai đồng nhất và phân bố chuẩn.
- Kiểm định Mann-Whitney; các giả định là các quan sát độc lập và phương sai đồng nhất.

Quyết định chọn kiểm định thống kê cuối cùng

Các giả định được kiểm tra như mục 4.7.

Số ngày nằm viện không có phân bố chuẩn và thỏa mãn phương sai đồng nhất.

Thực hiện kiểm định thống kê Mann-Whitney.



Viết báo cáo phương pháp

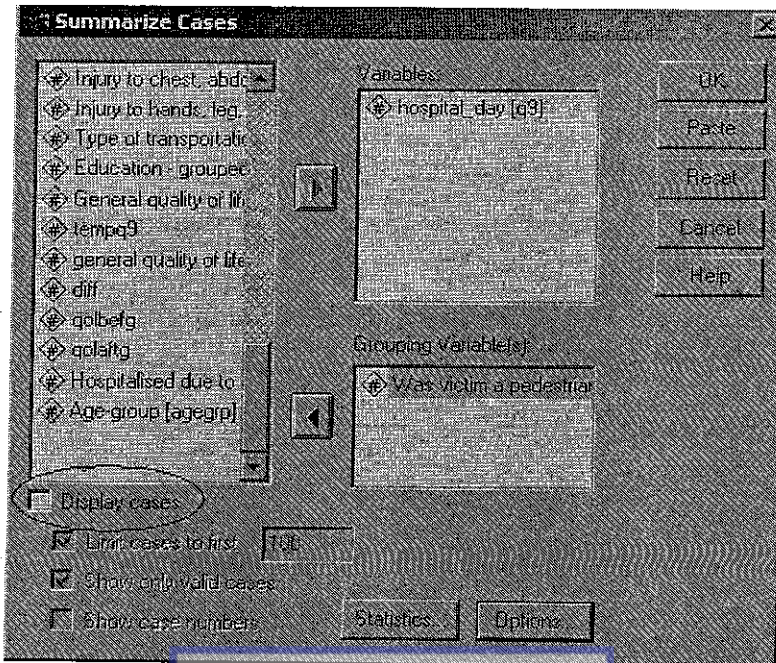
Báo cáo bạn có thể viết như sau:

Biến phụ thuộc là số ngày nằm viện được chứng minh là không có phân bố chuẩn nên để so sánh trung vị số ngày nằm viện của những người đi bộ và bị tai nạn giao thông với những người không đi bộ chúng ta sử dụng kiểm định Mann-Whitney.



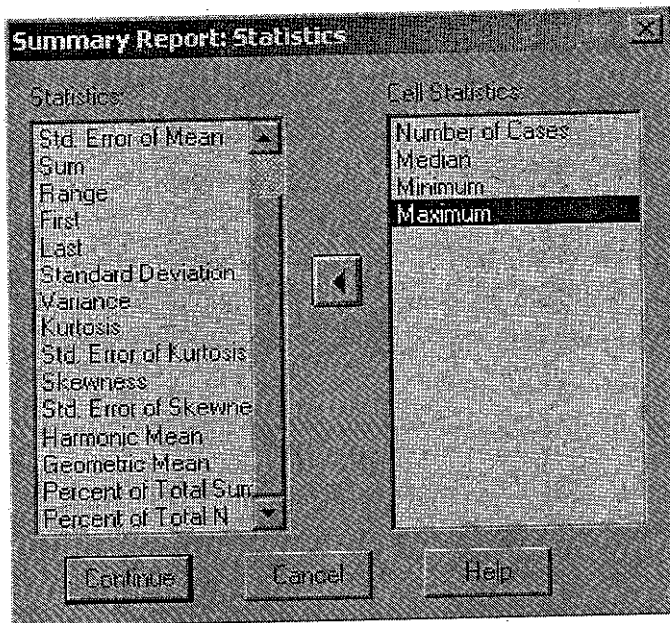
SỬ DỤNG SPSS ĐỂ KIỂM ĐỊNH GIẢ THUYẾT – Hai trung vị

1. Đầu tiên, bạn muốn có kết quả về thống kê mô tả cho mỗi nhóm có nghĩa là trung vị số ngày nằm viện của những người đi bộ thì bạn chọn từ thực đơn **Analyse/ Reports/Case Summaries**.
2. Từ danh sách các biến, đánh dấu vào biến phụ thuộc mà bạn muốn phân tích. Trong trường hợp này là q9 (số ngày nằm viện), và chuyển biến đó sang ô **Test Variable(s)** bằng cách kích vào mũi tên phía trên.
3. Từ danh sách các biến, đánh dấu vào biến độc lập mà bạn muốn sử dụng (có nghĩa là các nhóm mà bạn muốn so sánh) trong ví dụ này là pedestrn và chuyển biến đó sang ô **Grouping Variable** bằng cách kích vào mũi tên phía dưới.
4. Bởi vì trong trường hợp này báo cáo tóm tắt một nhóm không muốn chi tiết đến các thành viên nên bỏ chọn **Display Cases**.



5. Bây giờ kích vào Statistics. Bạn sẽ nhìn thấy một hộp thoại có dạng sau..

6. Chọn các đại lượng thống kê mô tả mà bạn muốn: trong ví dụ chúng ta chọn các đại lượng sau Number of cases, Median, Minimum & Maximum, và chuyển lần lượt các đại lượng đó vào ô Cell Statistics. Bây giờ kích Continue.



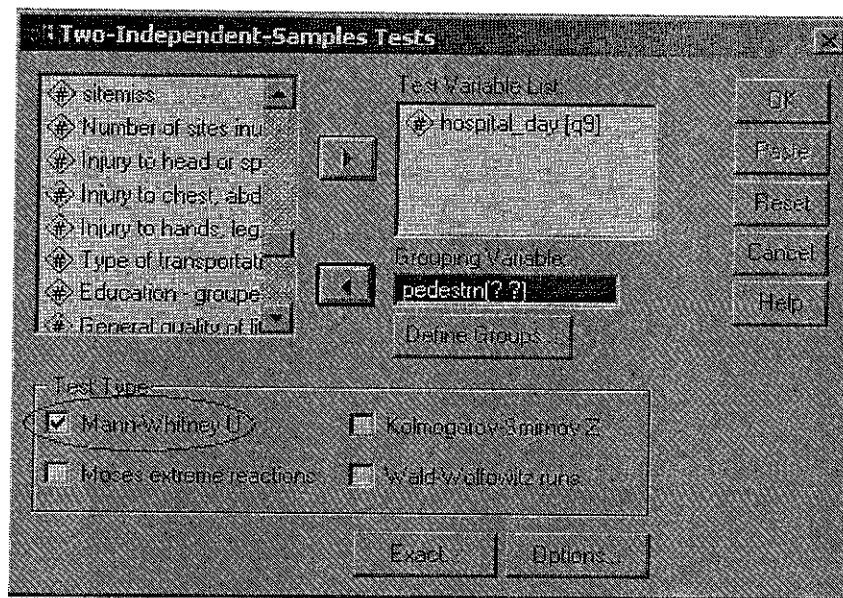
7. Bây giờ kích OK. Bạn sẽ nhận được được một kết quả dạng như sau và biến phụ thuộc được tóm tắt thành các nhóm. Tuy nhiên, bạn chưa thực hiện kiểm định thống kê để xem liệu các nhóm có thực sự khác biệt hay không.

8. Từ thực đơn chọn: **Analyse/Nonparametric Tests/2 Independent-Samples**, bạn sẽ có một hộp thoại như ở dưới.

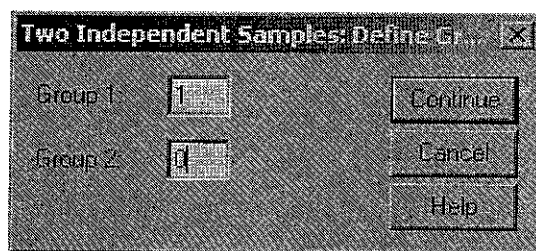
9. Đánh dấu vào ô chọn Mann-Whitney U trong phần Test Type để chọn kiểm định thống kê.

10. Từ danh sách các biến, đánh dấu vào biến phụ thuộc mà bạn muốn phân tích, trong ví dụ này là q9 (số ngày nằm viện), và chuyển biến đó sang ô Test Variable(s) bằng cách kích vào mũi tên phía trên.

11. Từ danh sách các biến, đánh dấu vào biến độc lập mà bạn muốn sử dụng (có nghĩa là các nhóm mà bạn muốn so sánh). Trong ví dụ này là pedestrn và kích vào mũi tên phía dưới để chuyển biến đó sang ô Grouping Variable.



12. Bây giờ bạn hãy định nghĩa các loại của biến độc lập để SPSS hiểu được cái gì mà bạn muốn so sánh (có nghĩa là nhóm nào là nhóm đi bộ nhóm nào là nhóm thể hiện không đi bộ). Để làm được điều đó, bạn đánh dấu vào biến độc lập (pedestrn) và kích vào Define Groups. Bạn sẽ thấy một hộp thoại hiện ra. Bạn cần phải chỉ ra cho SPSS biết mã đi bộ là gì, (ở đây mã của nhóm đi bộ là 1), mã không đi bộ là gì (ở đây dữ liệu mã nhóm không đi bộ là 0) bằng cách đánh mã vào các ô Group 1 và Group 2 tương ứng. Bây giờ kích vào Continue.



13. Bây giờ kích vào OK.

Kết quả đầu ra của bạn sẽ xuất hiện trong một cửa sổ riêng biệt – cửa sổ kết quả và có dạng như sau:

KẾT QUẢ

Summarize

Case Summaries

hospital_day				
Was victim a pedestrian?	N	Median	Minimum	Maximum
No	660	7.00	1	200
Yes	91	5.00	1	120
Total	751	7.00	1	200

Mann-Whitney Test

Ranks

hospital_day				
Was victim a pedestrian?	N	Mean Rank	Sum of Ranks	
No	660	381.73	251941.50	
Yes	91	334.45	30434.50	
Total	751			

Test Statistics^a

hospital_day	
Mann-Whitney U	26248.500
Wilcoxon W	30434.500
Z	-1.962
Asymp. Sig. (2-tailed)	.050

a. Grouping Variable: Was victim a pedestrian?

PHIÊN GIẢI

Gộp toàn bộ kết quả phân tích thống kê, báo cáo của bạn có thể viết dạng như sau:

Khi xem xét số ngày nằm viện của các nạn nhân, chúng ta có đủ bằng chứng để kết luận rằng có sự khác biệt về trung vị số ngày nằm viện của những người bị tai nạn khi đang đi bộ và không đi bộ ($Z = -1,96$, $n = 751$, $p = 0,05$). Trung vị số ngày nằm viện của những nạn nhân đi bộ thấp hơn hai ngày so với những người không đi bộ.

4.5.8. So sánh trung vị của ba hay nhiều hơn ba nhóm

LẬP KẾ HOẠCH PHÂN TÍCH – TRUNG VỊ CỦA HAI HAY NHIỀU HƠN

Xét giả thuyết sau:



H_0 : Trong những người không đi bộ, trung bình số ngày nằm viện giống nhau ở tất cả các loại phương tiện giao thông.

Lập kế hoạch phân tích cho kiểm định giả thuyết này có dạng như sau:

Mô tả các biến

Biến phụ thuộc: số ngày nằm viện; biến liên tục.

Biến độc lập là loại tai nạn, biến phân loại; 4 nhóm.

Mô tả các mối liên quan

Mối liên quan được tóm tắt qua trung bình và phương sai: trung bình, độ lệch chuẩn nếu biến có phân bố chuẩn, trung vị và khoảng nếu biến không có phân bố chuẩn.

Bảng giả

Loại tai nạn giao thông	số ngày nằm viện	
	Trung bình (mean)	Độ biến thiên (s.d.)
Ô tô		
Xe đạp		
Xe máy		
Khác		

Xác định các kiểm định thống kê có thể dùng

Sử dụng các kiểm định trong bảng 1(tr.193) là:

- Kiểm định ANOVA một chiều; các giả định: các quan sát độc lập, phân bố chuẩn và phương sai đồng nhất
- Kiểm định ANOVA Kruskal–Wallis; các giả định các quan sát độc lập và phương sai đồng nhất

Chọn kiểm định thống kê cuối cùng

Các giả định được kiểm tra theo từng mục 4.7.

Số ngày nằm viện không có phân bố chuẩn và thoả mãn phương sai đồng nhất.

Thực hiện kiểm định ANOVA Kruskal–Wallis.

Viết báo cáo phương pháp

Phần các phương pháp của bạn được viết có dạng sau; lưu ý rằng giả thuyết này chỉ xét với những người chấn thương giao thông không đi bộ cho nên trước khi phân tích bạn chỉ cần chọn những trường hợp không đi bộ trong bộ số liệu của bạn và có thể cập phần này trong các phương pháp của bạn.

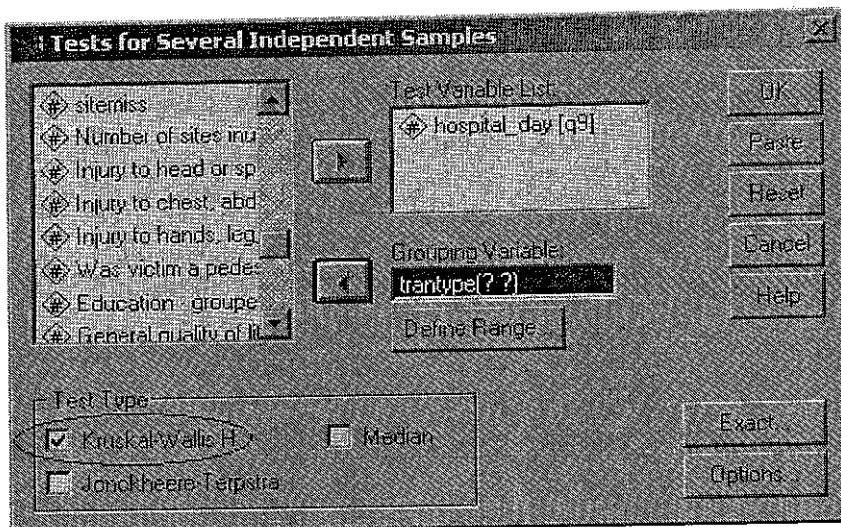
Do số ngày nằm viện của các nạn nhân không có phân bố chuẩn nên để so sánh trung vị của độ dài số ngày nằm viện của 4 nhóm tại nạn giao thông (loại trừ nhóm những người đi bộ) chúng ta sử dụng phân tích phương sai Kruskal-Wallis (hai phía).



SỬ DỤNG SPSS ĐỂ KIỂM ĐỊNH THỐNG KÊ – NHIỀU HƠN HAI TRUNG VỊ

Để tính trung vị của số ngày nằm viện của từng loại tai nạn giao thông bạn theo các bước được mô tả trong bài 3 phần 3.5.2, nếu SPSS không cho bạn kết quả đó trong phần đầu ra của kiểm định ANOVA Kruskal-Wallis. Bạn có thể tính các tóm tắt số liệu theo cách sau:

1. Chạy kiểm định ANOVA Kruskal-Wallis, từ menu chọn: **Analyse/Nonparametric Tests/K Independent Samples**. Bạn sẽ có một hộp thoại như dưới đây.

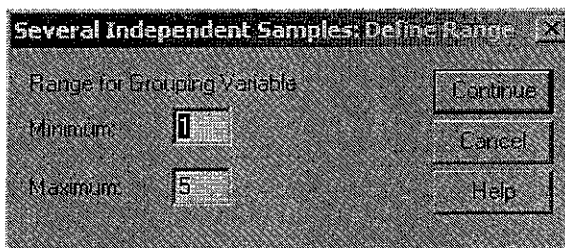


2. Từ danh sách các biến, đánh dấu vào biến phụ thuộc mà bạn muốn phân tích trong trường hợp này là q9 (số ngày nằm viện) và chuyển biến đó sang ô Test Variable List bằng cách kích vào mũi tên phía trên.

3. Từ danh sách các biến, đánh dấu vào biến độc lập mà bạn muốn sử dụng (có nghĩa là các nhóm mà bạn muốn so sánh). Trong ví dụ này là trantype và chuyển biến đó sang ô Grouping Variable bằng cách kích vào mũi tên phía dưới.

4. Hãy kiểm tra là bạn đã chọn Kruskal–Wallis H trong hộp Test Type chưa?

5. Bạn phải chỉ ra cho SPSS hiểu được khoảng số liệu của biến phụ thuộc (nhóm) có thể nhận, trong ví dụ này là trantype B được mã hoá là từ 1 đến 5. Để làm được điều này, đánh dấu vào trantype trong ô Grouping Variable và kích vào Define Range. Bạn sẽ có một hộp thoại dạng sau. Nhập giá trị lớn nhất và giá trị nhỏ nhất vào các ô và kích Continue.



6. Bây giờ kích OK.

Kết quả của bạn sẽ xuất hiện ở một cửa sổ riêng biệt – cửa sổ kết quả có dạng như sau .

KẾT QUẢ

Summarize

Report

hospital_day

Type of transportation in which victim travelling	N	Median	Minimum	Maximum
motorised vehicle	31	15.00	1	120
Bicycle	171	7.00	1	90
Motorsed bike	431	7.00	1	200
Other	27	7.00	1	60
Total	660	7.00	1	200

NPar Tests**Kruskal-Wallis Test**

Ranks

	Type of transportation in which victim travelling	N	Mean Rank
hospital_day	motorised vehicle	31	428.34
	Bicycle	171	297.15
	Motorised bike	431	335.30
	Other	27	352.76
	Total	660	

Test Statistics(a,b)

	hospital_day
Chi-Square	14.218
df	3
Asymp. Sig.	.003

a. Kruskal-Wallis Test

b. Grouping Variable: Type of transportation in which victim travelling

PHIÊN GIẢI

Trong trường hợp này, số ngày nằm viện của các loại tai nạn giao thông có sự khác biệt có ý nghĩa thống kê. Chúng ta có thể viết báo cáo sau:

Trung vị số ngày nằm viện của các loại tai nạn giao thông có sự khác biệt có ý nghĩa thống kê (Kruskal-Wallis test, $n = 660$, $p = 0,003$).

Kiểm định Kruskal-Wallis tương đương với giai đoạn đầu tiên của phân tích phương sai ở trên. Bạn nên sử dụng các kiểm định Mann-Whitney để thực hiện kiểm định ở giai đoạn 2 là kiểm tra từng cặp và chạy nhiều kiểm định để so sánh mỗi cặp – nhóm. Xem phần sử dụng SPSS để kiểm định thống kê – hai giá trị trung vị để biết chi tiết về cách chạy các phép so sánh trong SPSS.

4.5.9. Không nhóm – tất cả các biến là liên tục và chuẩn

LẬP KẾ HOẠCH PHÂN TÍCH – CẢ HAI BIẾN LIÊN TỤC (CẢ HAI ĐIỀU LÀ PHÂN BỐ CHUẨN)

Xét giả thuyết thống kê sau:

H_0 : Trung bình điểm chất lượng cuộc sống sau khi bị chấn thương không liên quan đến tuổi của người bị chấn thương.

Kế hoạch phân tích giả thuyết thống kê này sẽ có dạng như sau:



Mô tả các biến

Một biến phụ thuộc là điểm chất lượng cuộc sống, liên tục.

Một biến độc lập tuổi (tính bằng đơn vị là năm), liên tục.

Mô tả mối liên quan

Mối liên quan được thể hiện qua biểu đồ chấm điểm để xác định hướng

Xác định các loại kiểm định thống kê

Khi có nhiều giá trị (liên tục) chúng ta không sử dụng so sánh nhóm mà thay vào đó là mô tả mối quan hệ giữa hai nhóm. Một cách đơn giản nhất tóm tắt mối quan hệ thông qua một giá trị duy nhất là tính hệ số tương quan.

Sử dụng bảng 1(tr.193) ta có các kiểm định có thể sử dụng được như sau:

- Tương quan Pearson's; các giả định là các quan sát độc lập và cả hai biến đều có phân bố chuẩn.
- Tương quan hạng Spearman; các giả định là các quan sát độc lập (một hoặc cả hai không có phân bố chuẩn).

Chọn kiểm định thống kê cuối cùng

Các giả định được kiểm tra theo từng mục 4.7.

- Cả biến điểm chất lượng cuộc sống và tuổi đều có phân bố chuẩn; hai biến có mối quan hệ tuyến tính.
- Thực hiện việc tương quan Pearson (được ký hiệu là r).

Viết báo cáo phương pháp

Phần mô tả phương pháp bạn có thể viết như sau:

Vì cả hai biến điểm chất lượng cuộc sống và tuổi đều có phân bố chuẩn nên chúng ta dùng hệ số tương quan Pearson để tóm tắt mối quan hệ giữa hai biến.

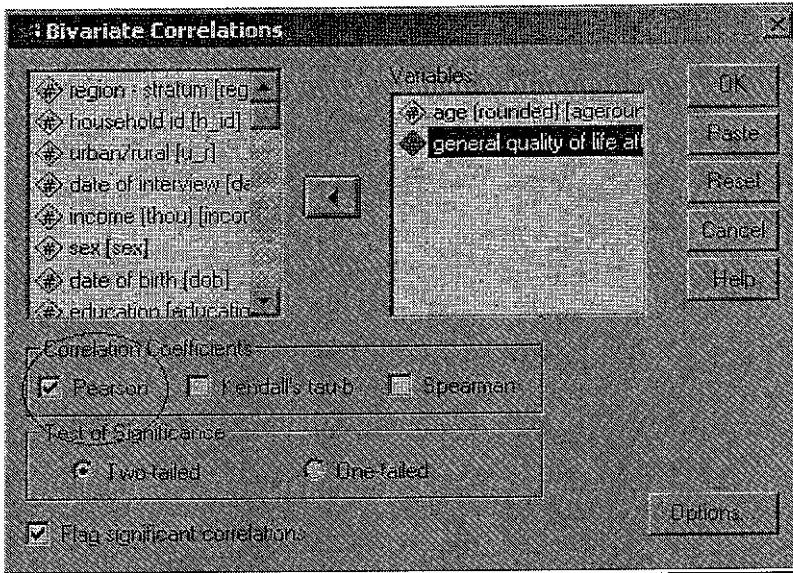
SỬ DỤNG SPSS ĐỂ KIỂM ĐỊNH GIẢ THUYẾT – Tương quan PEARSON

Để tìm hiểu mối quan hệ giữa hai biến này, bạn vẽ biểu đồ chấm điểm của hai biến. Trong ví dụ này là qol_ aft (điểm chất lượng cuộc sống sau khi bị chấn thương) và ageround (tuổi tính theo đơn vị năm). Bạn nên xem lại bài 3 để biết cách dùng SPSS để vẽ biểu đồ. Biểu đồ kết quả của bạn được tạo ra theo cách sau đây:

1. Từ thanh thực đơn chọn **Analyse/Correlate/Bivariate** để tính giá trị tương quan Pearson. Bạn sẽ thấy xuất hiện một cửa sổ như dưới đây.

2. Từ danh sách các biến đánh dấu vào từng biến mà bạn muốn phân tích. Trong ví dụ này là **qol_aft** (điểm chất lượng cuộc sống sau khi chấn thương) và **ageround** (tuổi tính bằng năm), và chuyển hai biến đó sang ô **Test Variable List** cùng một lúc bằng cách kích vào mũi tên.

3. Kích vào ô **Pearson** trong phần **Correlation Coefficients**.

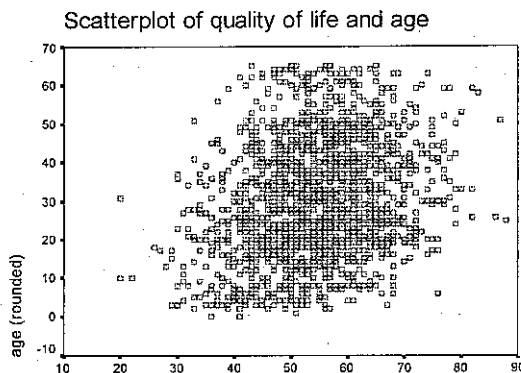


4. Bây giờ kích **OK**.

Kết quả của bạn sẽ xuất hiện ở cửa sổ kết quả như sau:

KẾT QUẢ

Biểu đồ



Tương quan

Correlations

		age (rounded)	general quality of life after injury
age (rounded)	Pearson Correlation	1	.236(**)
	Sig. (2-tailed)	.	.000
	N	1721	1693
general quality of life after injury	Pearson Correlation	.236(**)	1
	Sig. (2-tailed)	.000	.
	N	1693	1693

** Correlation is significant at the 0.01 level (2-tailed).

PHIÊN GIẢI

Hệ số tương quan nằm khoảng từ -1 đến $+1$, hệ số tương quan bằng 0 có nghĩa là không có mối quan hệ giữa hai biến; $+1$ có nghĩa là có mối quan hệ thuận và chặt (thấp điểm nhất thì tuổi thấp nhất, thấp điểm thứ hai thì tuổi thấp thứ hai..., cao điểm nhất thì tuổi cao nhất) và -1 có nghĩa là liên quan nghịch và chặt (tuổi thấp nhất thì có điểm cao nhất..., tuổi cao nhất thì có điểm thấp nhất). Lưu ý rằng, các hệ số tương quan chỉ tóm tắt độ lớn cho mối quan hệ tuyến tính. Bất cứ mối quan hệ nào khác không phải tuyến tính thì không được dùng cho nên nếu $r = 0$ có nghĩa là không có mối liên quan gì cả hoặc là mối liên quan nào đó có dạng phức tạp hơn quan hệ tuyến tính. Kết quả của ví dụ trên có thể tóm tắt như sau:

Mối tương quan giữa tuổi và điểm chất lượng cuộc sống sau khi chấn thương là yếu (Pearson's $r = 0,24$, $n = 1693$, $p < 0,001$).

Trong trường hợp này, lưu ý rằng, mặc dù giá trị p chỉ ra là mối quan hệ tuyến tính giữa tuổi và điểm chất lượng cuộc sống có ý nghĩa thống kê, nhưng do hệ số tương quan bằng $0,24$ nên có thể nói rằng mối quan hệ giữa tuổi và điểm chất lượng cuộc sống sau khi bị chấn thương là yếu. Việc có ý nghĩa thống kê trong trường hợp này có thể là do cỡ mẫu của nghiên cứu lớn. Đây là một ví dụ cho chúng ta thấy rằng trong nhiều trường hợp chúng ta phải xét đến cả giá trị thống kê chứ không chỉ dựa vào giá trị p để quyết định kết quả của kiểm định giả thuyết. Mặc dù một kiểm định giả thuyết chỉ ra là có ý nghĩa thống kê nhưng điều quan trọng là khi phiên giải kết quả ở đây phải dựa vào thực tế của nghiên cứu. Chúng ta sẽ thảo luận vấn đề này sâu hơn trong bài 5.

4.5.10. Không nhóm – cả hai biến là liên tục và có phân bố chuẩn

LẬP KẾ HOẠCH PHÂN TÍCH – HAI BIẾN LIÊN TỤC – Dạng câu hỏi khác

Xét giả thuyết thống kê từ phân trên (hai biến liên tục và có phân bố chuẩn):

H_0 : Trung bình điểm chất lượng cuộc sống không liên quan đến tuổi của nạn nhân bị chấn thương.

Một cách để tóm tắt mối quan hệ là sử dụng hệ số tương quan. Tuy nhiên, trường hợp đó chỉ có thể ước lượng được cho trường hợp mối liên quan giữa hai biến là tuyến tính. Trong một vài trường hợp chúng ta có thể chỉ ra trực tiếp mối quan hệ này hay có nghĩa là một biến phụ thuộc vào biến kia. Trong trường hợp đó, nếu chúng ta biết mối quan hệ trong các thành phần sẽ hữu ích cho chúng ta sẽ có thể dự đoán được giá trị biến phụ thuộc, trong ví dụ này là điểm chất lượng cuộc sống từ các giá trị đã biết của biến độc lập, trong ví dụ này là tuổi của nạn nhân. Điều này yêu cầu các loại kiểm định khác nhau như là hồi quy tuyến tính. Nếu câu hỏi nghiên cứu là liệu có thể dựa vào biến độc lập để dự báo biến phụ thuộc thì lập kế hoạch phân tích có dạng:

Mô tả các biến

Biến phụ thuộc là điểm chất lượng cuộc sống, biến liên tục.

Biến độc lập là tuổi (tính bằng đơn vị năm); biến liên tục.

Mô tả mối liên quan

Sử dụng biểu đồ chấm điểm để tóm tắt mối liên quan và xác định hướng, kỳ vọng là có mối quan hệ tuyến tính.

Xác định các kiểm định thống kê

Vì tất cả các giá trị là liên tục, nên hồi quy tuyến tính có thể được sử dụng.

Chọn kiểm định thống kê cuối cùng

Các giả định (độc lập, đồng nhất, quan hệ tuyến tính) được kiểm tra theo từng phần như trong mục 4.7.

Trên biểu đồ chấm điểm xuất hiện mối liên quan tuyến tính chắc chắn rằng không có một mối quan hệ gì phức tạp hơn mối quan hệ tuyến tính. Thực hiện phép hồi quy tuyến tính.



Viết báo cáo phương pháp

Phần các phương pháp của bạn nên viết có dạng sau:

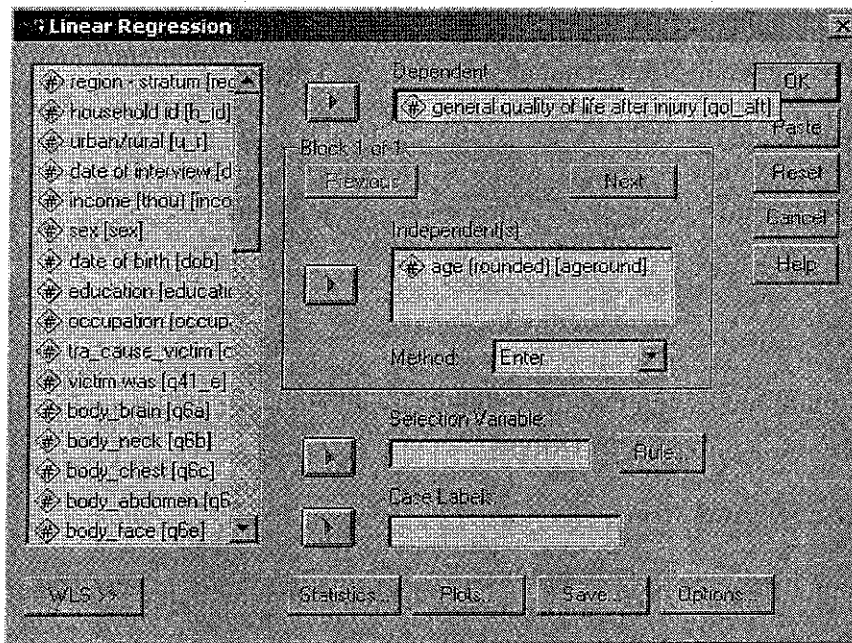
Chúng ta dùng hồi quy tuyến tính để miêu tả mối quan hệ giữa điểm chất lượng cuộc sống và tuổi.



DÙNG SPSS ĐỂ KIỂM ĐỊNH GIÁ THUYẾT – Hồi quy tuyến tính đơn giản

Bạn hãy dùng biểu đồ chấm điểm để thể hiện mối quan hệ giữa hai biến trong ví dụ này là qol_aft (Quality of Life score after injury điểm chất lượng cuộc sống sau khi bị chấn thương) và ageround (tuổi tính theo năm). Bạn có thể tham khảo bài 3 để biết cách sử dụng SPSS vẽ biểu đồ chấm điểm. Biểu đồ không được chỉ ra đây.

1. Để chạy hồi quy tuyến tính đơn giản, từ thanh thực đơn bạn chọn **Analyze/Regression/Linear**. Bạn sẽ thấy một hộp thoại xuất hiện như sau:

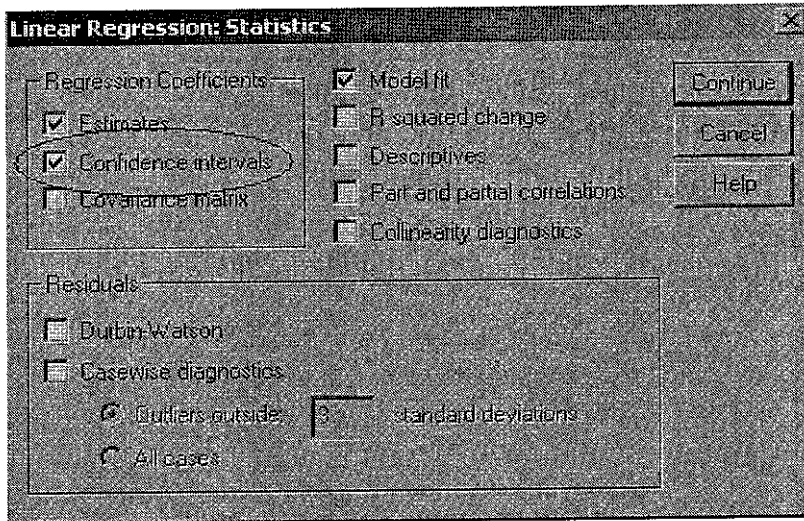


2. Từ danh sách các biến, đánh dấu vào biến phụ thuộc, trong ví dụ này là qol_aft và chuyển biến đó sang ô Dependent bằng cách sử dụng mũi tên.

3. Sau đó chọn biến độc lập, trong ví dụ này là ageround và dùng mũi tên để chuyển biến độc lập sang ô Independent(s).

4. Bây giờ kích vào Statistics. Bạn sẽ thấy một hộp thoại sau. Nếu bạn muốn SPSS tính khoảng tin cậy cho hệ số hồi quy, giá trị này được sử dụng để độ đo

mức chính xác của phép kiểm định, bạn chọn Confidence intervals trong hộp Regression Coefficients. Sau đó kích Continue.



5. Bây giờ kích OK.

KẾT QUẢ

Regression

Variables Entered/Removed(b)

Model	Variables Entered	Variables Removed	Method
1	age (rounded)(a)		Enter

a. All requested variables entered.

b. Dependent Variable: general quality of life after injury

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.236(a)	.056	.055	9.71852

a. Predictors: (Constant), age (rounded)

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9439.761	1	9439.761	99.946	.000(a)
	Residual	19714.350	1691	94.450		
	Total	169154.111	1692			

a. Predictors: (Constant), age (rounded)

b. Dependent Variable: general quality of life after injury



Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	50.015	.522		95.854	.000	48.992	51.039
	age (rounded)	.155	.018	.235	9.997	.000	.125	.186

a. Dependent Variable: general quality of life after injury

PHIÊN GIẢI

Kết quả ở trên có thể được phiên giải như sau:

Chất lượng cuộc sống có mối liên quan ý nghĩa với tuổi ($F_{1,1691} = 99,9, p < 0,001$). Đối với mỗi một tuổi tăng lên điểm chất lượng cuộc sống sẽ tăng 0,16 đơn vị (khoảng tin cậy 95% là 0,1; 0,19). Tuổi lý giải được 6% sự biến thiên của điểm chất lượng cuộc sống.

4.5.11. Không nhóm – cả hai biến liên tục nhưng không có phân bố chuẩn

LẬP KẾ HOẠCH PHÂN TÍCH – HAI BIẾN LIÊN TỤC (CẢ HAI HOẶC ÍT NHẤT MỘT BIẾN KHÔNG CÓ PHÂN BỐ CHUẨN)

Xét giả thuyết thống kê sau:

H_0 : trung bình điểm chất lượng cuộc sống sau khi chấn thương không liên quan đến số ngày nằm viện của bệnh nhân.

Kế hoạch phân tích cho kiểm định giả thuyết thống kê này có dạng như sau:

Mô tả các biến

Biến phụ thuộc là điểm chất lượng cuộc sống, biến liên tục.

Biến độc lập là số ngày nằm viện, biến liên tục.

Mô tả mối quan hệ

Dùng biểu đồ chấm điểm để mô tả mối quan hệ giữa hai biến để xác định hướng.

Xác định các kiểm định thống kê

Các kiểm định trong bảng 1(tr.193) có thể dùng là:

- Tương quan Pearson; các giả định là các quan sát độc lập, mối quan hệ giữa hai biến là tuyến tính và phân bố của hai biến là phân bố chuẩn.

- Tương quan hạng Spearman; các giả định là các quan sát độc lập, mỗi quan hệ giữa hai biến là quan hệ tuyến tính; một hoặc cả hai biến không có phân bố chuẩn).

Chọn loại kiểm định thống kê cuối cùng

Các giả định được kiểm tra theo từng mục 4.7.

Mặc dù điểm chất lượng cuộc sống có phân bố chuẩn nhưng số ngày nằm viện của nạn nhân lại không có phân bố chuẩn; mỗi quan hệ giữa hai biến là quan hệ tuyến tính.

Thực hiện kiểm định tương quan hạng Spearman.

Viết báo cáo phương pháp

Phần các phương pháp của bạn sẽ được viết dạng như sau:

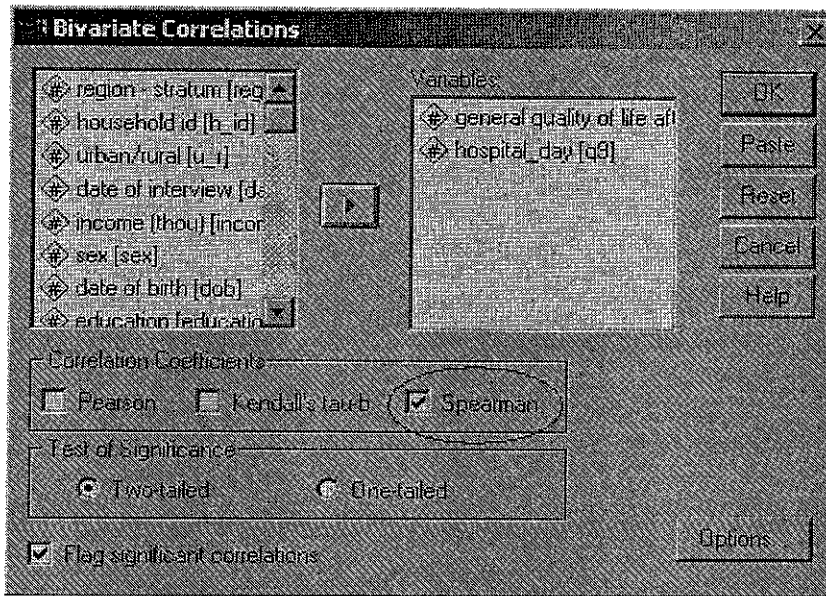
Do số ngày nằm viện của nạn nhân không có phân bố chuẩn nên chúng ta sử dụng hệ số tương quan hạng Spearman để tóm tắt mối quan hệ giữa điểm chất lượng cuộc sống và số ngày nằm viện.



SỬ DỤNG SPSS ĐỂ KIỂM ĐỊNH GIÁ THUYẾT – Tương quan SPEARMANS

Dùng biểu đồ chấm điểm để mô tả mối quan hệ giữa hai biến, trong ví dụ này là qol_aft (Quality of Life score after injury điểm chất lượng cuộc sống sau khi chấn thương) và q9 (số ngày nằm viện). Bạn nên tham khảo bài 3 để biết cách dùng SPSS để vẽ biểu đồ.

1. Để tính giá trị tương quan Spearmans, chọn thực đơn **Analyse/Correlate/Bivariate**. Bạn sẽ thấy một hộp thoại dạng như dưới đây:
2. Từ danh sách các biến, đánh dấu vào mỗi biến mà bạn muốn phân tích, trong ví dụ này là qol_aft và q9 sau đó chuyển đồng thời hai biến này sang ô Test Variable List bằng cách kích vào dấu mũi tên.
3. Kiểm tra lại là bạn đã chọn kiểm định Spearman trong ô Correlation Coefficients.

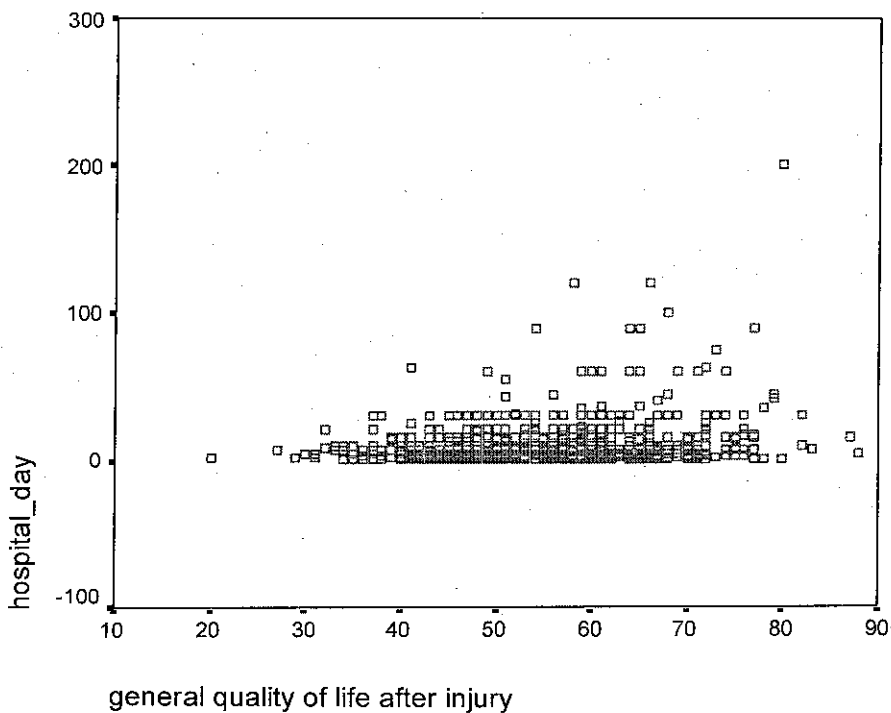


4. Bây giờ kích vào OK.

KẾT QUẢ

Graph

Scatterplot of length of hospital stay and quality of life after injury



Nonparametric Correlations

Correlations

			general quality of life after injury	hospital_day
Spearman's rho	general quality of life after injury	Correlation Coefficient	1.000	.090(*)
		Sig. (2-tailed)		.010
		N	1693	802
	hospital_day	Correlation Coefficient	.090(*)	1.000
		Sig. (2-tailed)	.010	
		N	802	810

*. Correlation is significant at the 0.05 level (2-tailed).

PHIÊN GIẢI

Tương quan Spearman sẽ giống với các phép tính như tương quan Pearson ngoài trừ việc chúng ta sử dụng thứ hạng của số liệu thay vì bản thân số liệu. Kết quả được phiên giải tương tự như phiên giải kết quả của tương quan Pearson. Hệ số tương quan nằm trong khoảng từ -1 đến $+1$, với $+1$ có nghĩa là mối quan hệ tương quan thuận và chặt (tuổi thấp nhất thì có điểm thấp nhất, tuổi thấp thứ 2 nếu điểm thấp thứ 2..., tuổi cao nhất thì có điểm cao nhất), và -1 có nghĩa là mối quan hệ nghịch và chặt (nhóm có tuổi thấp nhất có điểm cao nhất..., tuổi cao nhất có điểm thấp nhất). $r = 0$ có nghĩa “không có một liên quan gì” hoặc “một mối quan hệ gì đó phức tạp hơn tuyến tính”.

Cho nên kết quả phân tích trên có thể được phiên giải dạng như sau:

Số ngày nằm viện quan hệ nghịch với điểm chất lượng cuộc sống sau khi chấn thương (Spearman's $r = 0,09$, $n = 802$, $p = 0,010$).

Mặc dù, giá trị p chỉ ra có một mối liên quan có ý nghĩa thống kê giữa số ngày nằm viện và điểm chất lượng cuộc sống sau khi bị chấn thương nhưng do hệ số tương quan chỉ là $0,09$ nên mối quan hệ này có thể bỏ qua được. Các tính toán bao gồm trong cả hệ số tương quan Pearson và Spearman có độ nhạy cao khi mẫu lớn nên đôi khi mối quan hệ rất yếu (r rất gần 0) nhưng vẫn có ý nghĩa thống kê. Đây là một ví dụ điển hình cho việc việc có ý nghĩa thống kê nhưng không thoả mãn có ý nghĩa trong thực tế.

4.5.12. So sánh một tỷ lệ mẫu với một tỷ lệ lý thuyết hoặc quần thể

LẬP KẾ HOẠCH PHÂN TÍCH – MỘT TỶ LỆ

Xét giả thuyết thống kê sau.

H_0 : tỷ lệ chấn thương ở xương sống cột sống và ở đầu là 37%.

Kế hoạch phân tích của bạn bao gồm những bước sau:

Mô tả các biến

Biến phụ thuộc là chấn thương ở đầu và xương sống cột sống, nhị thức.

Không có biến độc lập, chỉ là một giá trị quần thể.

Mô tả mối tương quan

Vì biến phụ thuộc là phân loại nên mô tả bằng số đếm và thể hiện phần trăm.

Không cần bảng tóm tắt.

Xác định các kiểm định thống kê

Sử dụng các kiểm định trong bảng 2(tr.194) là kiểm định Khi bình phương một mẫu.

Chọn kiểm định thống kê cuối cùng.

Các giả định được kiểm tra theo từng mục 4.7.

Các đơn vị quan sát độc lập.

Thực hiện kiểm định khi bình phương một mẫu.

Viết báo cáo phương pháp

Phân các phương pháp của bạn cho kiểm định thống kê này có dạng sau:

Chúng ta sử dụng kiểm định khi bình phương một mẫu để so sánh tỷ lệ chấn thương ở đầu/xương sống, cột sống trong năm 2001 và tỷ lệ chấn thương năm 1997.



SỬ DỤNG SPSS ĐỂ KIỂM ĐỊNH THỐNG KÊ – KIỂM ĐỊNH KHI BÌNH PHƯƠNG

1. Từ thanh thực đơn chọn: **Analyse/Nonparametric Tests/Chi-Square**.
Bạn sẽ có một hộp thoại giống sau:

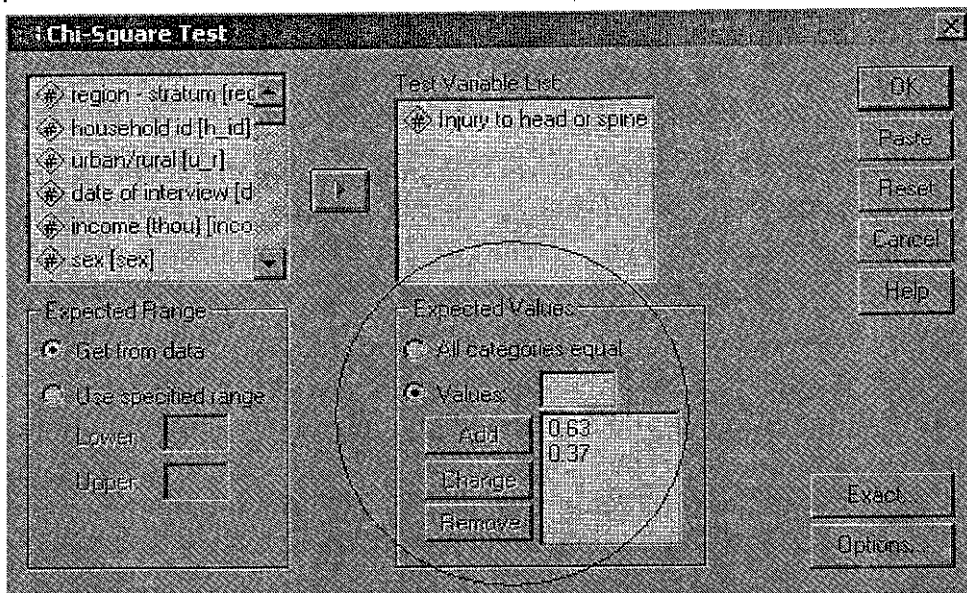
2. Từ danh sách các biến, đánh dấu biến mà bạn muốn phân tích, trong ví dụ này là headspin (chấn thương xương sống/cột sống hoặc ở đầu) và chuyển biến đó sang ô Test Variable List bằng cách kích vào mũi tên.

3. Bây giờ bạn phải chỉ cho SPSS biết tỷ lệ nào mà bạn mong muốn trên cơ sở giá trị quần thể mà bạn muốn sử dụng. Để thực hiện điều này bạn phải đưa giá trị vào ô Expected Values. Giá trị này phải nhỏ hơn 1. Trong ví dụ này giá trị kỳ vọng là 0,37 của tất cả các chấn thương giao thông bao gồm chấn thương đầu và chấn thương xương sống cột sống. Cho nên, chúng ta mong 0,63 tất cả các chấn thương không bao gồm chấn thương ở đầu và chấn thương cột sống.

4. Thêm tỷ lệ này vào ô Expected Values, bạn nhập số trong ô nhỏ bên cạnh vào từ Values, sau đó kích vào Add và giá trị này sẽ được chuyển sang ô lớn phía dưới.

Lưu ý: bạn phải nhập tỷ lệ kỳ vọng tương ứng với các giá trị theo đúng các giá trị đã được mã. Ví dụ biến headspin đã được mã là 0 nếu không bị chấn thương ở đầu/cột sống; 1 nếu chấn thương ở đầu/cột sống. Do đó tỷ lệ kỳ vọng phải được nhập là 0,63 là giá trị đầu tiên và 0,37 là giá trị thứ hai. Nếu bạn nhập chúng vào theo chiều ngược lại thì bạn sẽ tiến hành kiểm định giả thuyết sau:

H_0 : Tỷ lệ chấn thương ở đầu/cột sống là 63%. Điều này sẽ dẫn bạn đến một kết luận sai.



5. Kích OK.

KẾT QUẢ



NPar Tests

Chi-Square Test

Frequencies

Injury to head or spine

	Observed N	Expected N	Residual
Not injured at these sites	1089	1079.2	9.8
Injured	624	633.8	-9.8
Total	1713		

Test Statistics

	Injury to head or spine
Chi-Square(a)	.241
df	1
Asym. Sig.	.623

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 633.8.

PHIÊN GIẢI

Kiểm định Khi bình phương một mẫu (được biết là kiểm định khi bình phương cho tính phù hợp) tính toán sự khác nhau giữa giá trị quan sát và giá trị kỳ vọng cho mỗi ô trong bảng và gia quyền theo giá trị kỳ vọng. Giá thống kê là 0,623 và không có ý nghĩa thống kê tại mức ý nghĩa 5% ($p < 0,05$). Cho nên, kết quả phân tích này sẽ được trình bày như sau:

Tỷ lệ chấn thương ở đầu/cột sống xấp xỉ 36% trong nghiên cứu chấn thương giao thông năm 2001. Tỷ lệ này cũng tương tự như tỷ lệ báo cáo chấn thương giao thông quốc gia năm 1997, trong đó 37% chấn thương ở đầu/cột sống ($\chi^2_1 = 0,2, p = 0,623$).

Cũng lưu ý rằng giá trị kỳ vọng của các ô lớn hơn 5. Các giả định cho kiểm định khi bình phương chỉ chấp nhận nếu tất cả giá trị kỳ vọng của các ô đều lớn hơn 5 (chú ý, giá trị kỳ vọng chứ không phải giá trị quan sát, vì giá trị quan sát có thể nhỏ tới không). Điều này cũng không phải lúc nào cũng đúng vì trong thực tế kiểm định khi bình phương có thể có giá trị khi giá trị kỳ vọng nhận giá trị nhỏ tới mức bằng 2 (với một điều là không có nhiều ô nhỏ hơn 5).

Kiểm định khi bình phương cho tính phù hợp cũng có thể áp dụng cho biến phân loại mà có nhiều hơn hai loại. Ví dụ: tỷ lệ (tỷ lệ phần trăm) số nạn nhân tai nạn giao thông mà có điểm chất lượng sau khi chấn thương là kém, trung bình, và tốt. Rất hiếm khi tìm thấy hai kết quả khác nhau phụ thuộc vào biến đầu ra được phân loại như thế nào, như ví dụ này là điểm chất lượng cuộc sống.

Cần phải biết rằng tất cả các phân tích của bạn sẽ nhạy cảm với cách mà bạn phân loại biến đầu ra. Bạn phải nắm rõ các các phân loại trước khi bạn bắt đầu phân tích. Sẽ là không phù hợp nếu như bạn cố gắng tìm ra một điểm nào đó để tạo ra năng khả năng kiểm định thống kê có ý nghĩa. Do vậy bạn nên định nghĩa và định nghĩa lại các phân loại trong kế hoạch phân tích của mình.

4.5.13. So sánh tỷ lệ của hai nhóm

LẬP KẾ HOẠCH PHÂN TÍCH – SO SÁNH HAI TỶ LỆ

H_0 : Tỷ lệ chấn thương đầu và cột sống của nhóm đi xe tương đương với nhóm đi bộ hoặc ít hơn.

Kế hoạch phân tích bao gồm các bước sau:

Mô tả các biến

Một biến phụ thuộc là chấn thương ở đầu/cột sống, nhị phân.

Một biến độc lập là đi bộ, nhị phân.

Mô tả mối liên quan

Mối liên quan được tóm tắt theo dạng số đếm và tỷ lệ phần trăm.

Bảng giả.

	Đi bộ	Không đi bộ
Chấn thương ở đầu/ cột sống	n (%)	n (%)
Chấn thương không phải ở đầu/cột sống	n (%)	n (%)
Tổng	n (%)	n (%)

Xác định các kiểm định thống kê

Sử dụng các kiểm định thống kê trong bảng 2(tr.194) là kiểm định khi bình phương.

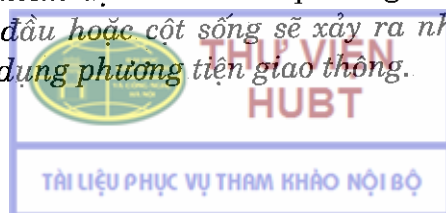
Chọn kiểm định thống kê cuối cùng.

Các giả định là các quan sát độc lập đã được kiểm tra. thực hiện kiểm định khi bình phương.

Viết báo cáo phương pháp

Phần các phương pháp trong kiểm định thống kê này sẽ có dạng như sau:

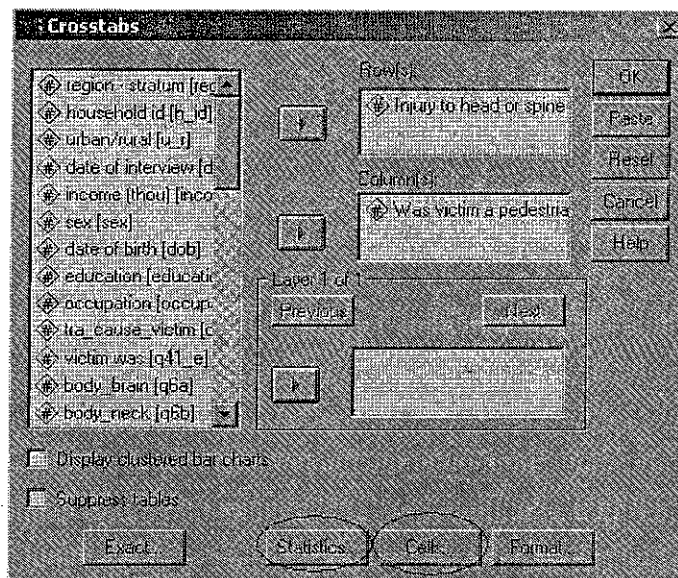
Chúng ta sử dụng kiểm định khi bình phương (một phía) để kiểm tra giả thuyết là chấn thương đầu hoặc cột sống sẽ xảy ra nhiều ở những người đi bộ hơn là những người sử dụng phương tiện giao thông.



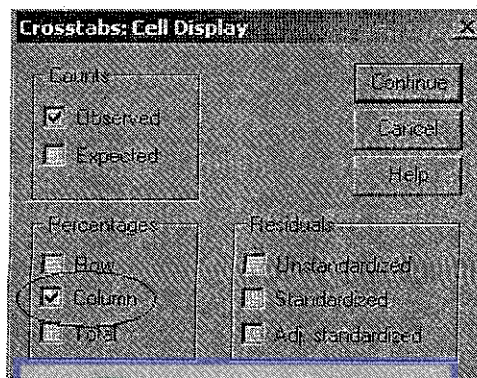


SỬ DỤNG SPSS ĐỂ KIỂM ĐỊNH THỐNG KÊ – Kiểm định khi bình phương

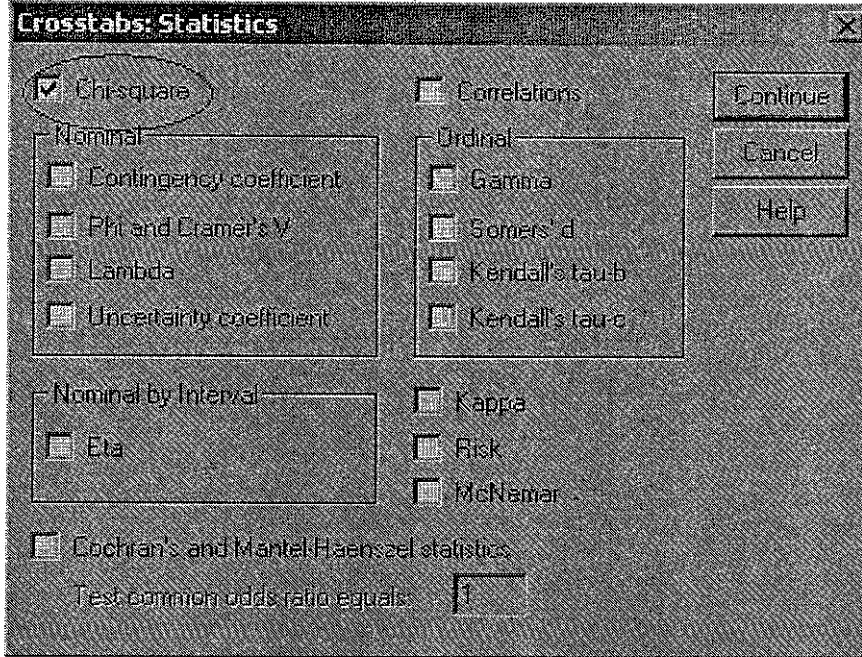
1. Từ thanh thực đơn chọn: **Analyse/Descriptive Statistics/Crosstabs**. một hộp thoại sẽ hiện ra.
2. Từ danh sách các biến, đánh dấu vào biến phụ thuộc mà bạn muốn phân tích. Trong ví dụ này là **headspin** (chấn thương ở đầu/ cột sống), và kích vào mũi tên để chuyển biến đó sang ô **Row(s)**.
3. Từ danh sách các biến, đánh dấu vào biến độc lập mà bạn muốn phân tích, trong ví dụ này là **pedestrn** (đi bộ hoặc không đi bộ), và kích vào mũi tên để chuyển sang ô **Column(s)**.



4. Nếu hiện thi tỷ lệ phần trăm cột những người đi bộ mà bị chấn thương đầu, bạn kích vào **Cells** và một hộp thoại mới xuất hiện dạng như sau. Đánh dấu vào ô cột và kích **Continue**.



5. Để tính kiểm định khi bình phương bằng SPSS bạn kích vào Statistics và một hộp thoại mới sẽ xuất hiện dạng sau. chọn Chi-square và kích tiếp vào Continue.



6. Bây giờ kích OK.

Kết quả của bạn sẽ như sau

KẾT QUẢ

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Injury to head or spine * Was victim a pedestrian?	1525	88.6%	196	11.4%	1721	100.0%

Injury to head or spine * Was victim a pedestrian? Crosstabulation



			Was victim a pedestrian?		Total
			No	Yes	
Injury to head or spine	Not injured at these sites	Count	861	94	955
		% within Was victim a pedestrian?	63.4%	56.0%	62.6%
	Injured	Count	436	74	570
		% within Was victim a pedestrian?	38.6%	44.0%	37.4%
Total		Count	1307	168	1525
		% within Was victim a pedestrian?	100.0%	100.0%	100.0%

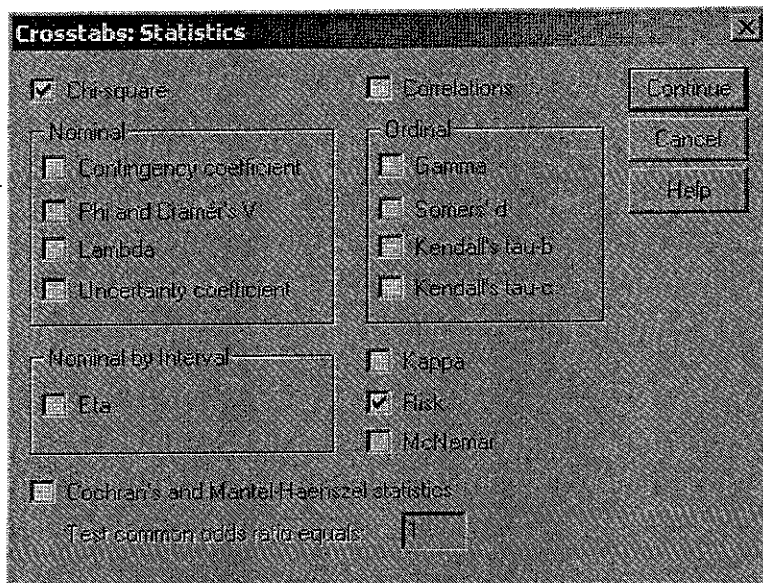
Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.589(b)	1	.058		
Continuity Correction(a)	3.276	1	.070		
Likelihood Ratio	3.526	1	.060		
Fisher's Exact Test				.063	.036
Linear-by-Linear Association	3.587	1	.058		
N of Valid Cases	1525				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 62.79.

Thông thường chúng ta sử dụng tỷ suất chênh và khoảng tin cậy để phiên giải kết quả. Tuy nhiên bạn cũng có thể đánh dấu vào ô 'Risk' trong màn hình chọn các giá trị thống kê cho ước lượng nguy cơ:



Bạn sẽ nhận được kết quả là :

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Injury to head or spine (Not injured at these sites / Injured)	1.367	.988	1.889
For cohort Was victim a pedestrian? = No	1.036	.997	1.076
For cohort Was victim a pedestrian? = Yes	.758	.569	1.010
N of Valid Cases	1525		

PHIÊN GIẢI

Có khoảng 44% những người đi bộ bị chấn thương đầu/cột sống so với 37% những người dùng phương tiện giao thông. Tỷ suất chênh của chấn thương đầu/cột sống trong những người đi bộ cao hơn 1,37 lần so với những người dùng phương tiện giao thông (vì số liệu này không được thu thập qua nghiên cứu thuần tập nên việc ước sử dụng hai nguy cơ khác là không chính xác) khoảng tin cậy bao gồm giá trị 1 chỉ ra rằng ước lượng này là chính xác:

Điều này chỉ ra rằng có sự khác biệt giữa các loại nạn nhân chấn thương nhưng lại không chỉ ra cụ thể là khác biệt cái gì?

Mỗi một lần tính giá trị kiểm định khi bình phương, bạn có thể có các kết quả hơi khác nhau. Điều này do trong kiểm định khi bình phương giá trị p chỉ là giá trị xấp xỉ, ngoài ra chúng ta còn có kết quả của một vài phương pháp khác như Likelihood, Pearsons. Một vài phần mềm thống kê có thể tính giá trị p chính xác khi thích hợp (kiểm định chính xác Fisher, không phải là một kiểm định khi bình phương), mặc dù đây là một phép tính đòi hỏi nhiều tính toán. Và như vậy, các kiểm định xấp xỉ khi bình phương là cần thiết. Kết quả trên đây cho phép chúng ta chọn lựa một trong 3 kiểm định thống kê khác nhau và đồng thời cũng cho giá trị xác suất chính xác.

Cũng giống như kiểm định khi bình phương cho một mẫu chúng ta cũng phải chú ý rằng giá trị kì vọng của ô phải lớn hơn 5. Một kiểm định khi bình phương có giá trị là tất cả các giá trị kì vọng của ô phải lớn hơn 5 (lưu ý: giá trị kì vọng chứ không phải giá trị quan sát, giá trị quan sát có thể bằng không). Tuy nhiên quy ước này cũng mang tính chất hơi bảo thủ, trên thực tế kiểm định khi bình phương có thể kiểm định đúng khi số thậm chí khi giá trị kì vọng của một ô nào đó nhận giá trị bằng 2 (không nhiều quá các ô trên có số nhỏ hơn 5).



SPSS sẽ chỉ ra là có bất kỳ một ô nào nhỏ hơn 5 nhưng vẫn thực hiện kiểm định khi bình phương cho bạn. Trong trường hợp này, khi phân giải kết quả bạn nên cẩn thận để tránh đưa ra các kết luận sai.

Kiểm định khi bình phương được trình bày nhiều nhất trong các tài liệu thống kê là kiểm định khi bình phương Pearson. Tuy nhiên, khi bảng chỉ có 2 hàng và 2 cột thì chúng ta nên áp dụng hiệu chỉnh liên tục cho công thức Pearson. Như vậy kiểm định thống kê chính xác nhất cho kết quả trên sẽ là kiểm định khi bình phương có hiệu chỉnh liên tục. Kiểm định thống kê Linear-by-linear chỉ phù hợp khi một hoặc cả hai biến của chúng ta là biến thứ bậc và có ít nhất 3 loại. Trong trường hợp này máy tính cả giá trị thống kê chính xác cho nên chúng ta có thể chọn giá trị này. Nó cũng tương đương với giá trị hiệu chỉnh liên tục. Tuy nhiên, phần này thảo luận về kiểm định khi bình phương nên chúng ta sẽ chọn giá trị kết quả của hiệu chỉnh liên tục để đưa vào báo cáo:

Có sự khác biệt giữa tỷ lệ chấn thương đầu/cột sống ở những người đi bộ so với những người dùng phương tiện giao thông. Có 44% những người bị chấn thương đầu/cột sống ở những người đi bộ nhưng chỉ có 37% những người bị chấn thương loại này khi dùng phương tiện giao thông. Tỷ suất chênh chỉ ra sự khác nhau của hai tỷ lệ này là 1,37 (khoảng tin cậy 95% 0,99 – 1,89). Mặc dù sự khác nhau này có ý nghĩa trong y tế công cộng nhưng chúng ta lại không đủ bằng chứng để kết luận rằng sự khác nhau giữa hai nhóm là có ý nghĩa thống kê ($\chi^2_1 = 3,3, p = 0,070$).

4.5.14. So sánh tỷ lệ của ba hay nhiều hơn ba nhóm

LẬP KẾ HOẠCH PHÂN TÍCH – SO SÁNH TỶ LỆ NHIỀU HƠN HAI TỶ LỆ

H₀: Tỷ lệ những người nhận được điểm chất lượng cuộc sống thấp là giống nhau không kể đến mức độ chấn thương, được đo dựa trên vị trí bị chấn thương nặng nhất khi va chạm.

Kế hoạch phân tích bao gồm các phần sau:

Mô tả các biến

Biến phụ thuộc là điểm chất lượng thấp, dạng đo lường nhị phân.

Biến độc lập: vị trí chấn thương; phân loại; 3 nhóm.

Mô tả mối liên quan

Tỷ lệ phần trăm và số lượng là mô tả của mối liên quan.

Xác định các loại kiểm định thống kê

Sử dụng bảng 2(tr.194) để chọn kiểm định thống kê

Chọn kiểm định thống kê cuối cùng

Các giả định được kiểm tra như từng mục 4.7.

Giả định các đơn vị quan sát độc lập thoả mãn, thực hiện kiểm định khi bình phương.

Viết báo cáo phương pháp

Phần các phương pháp của bạn cho kiểm định thống kê này có thể viết dạng như sau:

Chúng ta sử dụng kiểm định khi bình phương để so sánh tỷ lệ các nạn nhân chấn thương có điểm chất lượng cuộc sống thấp qua các mức độ chấn thương. Các mức độ chấn thương được đo bằng vị trí chấn thương.



SỬ DỤNG SPSS ĐỂ KIỂM ĐỊNH GIẢ THUYẾT – KIỂM ĐỊNH KHI BÌNH PHƯƠNG

Thực hiện kiểm định khi bình phương trong SPSS như các bước trong phần 4.1.13 ở trên.

KẾT QUẢ

Most severe injury * QOL score after injury - categorised Crosstabulation

			QOL score after injury - categorised		Total
			Suboptimal QOL	Adequate QOL	
Most severe injury	head-spine	Count	216	404	620
		% within Most severe injury	34.8%	65.2%	100.0%
	torso	Count	131	241	372
		% within Most severe injury	35.2%	64.8%	100.0%
	limbs	Count	204	443	647
		% within Most severe injury	31.5%	68.5%	100.0%
Total	Count	551	1088	1639	
	% within Most severe injury	33.6%	66.4%	100.0%	



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2.103 ^a	2	.349
Likelihood Ratio	2.111	2	.348
Linear-by-Linear Association	1.571	1	.210
N of Valid Cases	1639		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 125.06.

PHIÊN GIẢI

Kiểm định khi bình phương để so sánh nhiều hơn hai tỷ lệ chính là tính toán so sánh chỉ có hai tỷ lệ. Trong trường hợp này, số lượng quan sát và kỳ vọng rất giống nhau và kiểm định không có ý nghĩa thống kê ở mức 5%. Lưu ý rằng vì bảng này không phải bảng 2 x 2 nên không có hiệu chỉnh liên tục. Kết quả có thể viết dạng như:

Có tất cả 66,4% những người có điểm chất lượng cuộc sống thấp. Chúng ta không có đủ bằng chứng để chỉ ra rằng tỷ lệ này khác nhau theo vị trí chẩn thương ($\chi^2_2 = 2,1, p = 0,349$).

Các phép tính và phiên giải cũng tương tự khi biến phụ thuộc có nhiều hơn hai loại.

4.5.15. Mối liên quan của biến phụ thuộc phân loại với biến liên tục

Trong các mục chúng ta đã xem xét các giả thuyết giữa một biến phụ thuộc liên tục và một biến độc lập phân loại. Những kiểm định được sử dụng là:

Mối liên quan	Kiểm định
Hai giá trị trung bình	t không ghép cặp
Hai giá trị trung vị	Mann-Whitney
Nhiều giá trị trung bình	ANOVA
Nhiều giá trị trung vị	Kruskal-Wallis ANOVA

Nếu có một biến phụ thuộc phân loại và một biến độc lập liên tục, như một trong những dạng trên nhưng với trật tự đảo ngược được xem xét trong mục 4.5.2 và 4.5.3, chúng ta có thể sử dụng các kiểm định giống như vậy để xem xét mối liên quan của chúng. Tất cả các kiểm định thống kê này chứng minh một sự kết hợp (không có hướng). Kết quả của các kiểm định này sẽ giống như ở các phần trên và nên được giải thích chính xác theo cùng một cách. Tuy nhiên khi phiên giải kết quả bạn cần phải nhớ đâu là biến độc lập và đâu là biến phụ thuộc.

4.6. Trình bày kết quả của các phân tích suy luận

Một trong những giá trị của việc viết báo cáo trong kế hoạch phân tích là chúng ta có thể sử dụng chúng trong báo cáo cuối cùng của chúng ta. Bạn sẽ có đủ thông tin để viết phần phương pháp phân tích và những ý tưởng hay về những gì sẽ đề cập đến trong phần kết quả nghiên cứu.

Trong điều tra về chấn thương giao thông của quốc gia, phần kế hoạch phân tích và kết quả thu được của cuộc điều tra được trình bày tóm tắt dưới đây. Đây là một ví dụ gợi ý cho bạn cách viết một báo cáo cho những phân tích tương tự tuy nhiên mỗi người đều sẽ có những phong cách riêng của mình.

Một ví dụ về viết báo cáo.

Phương pháp phân tích:

Điểm của chất lượng cuộc sống (QoL) trước thời điểm chấn thương là phân bố chuẩn vì vậy kiểm định t một mẫu đã được dùng để so sánh giữa quần thể điều tra với quần thể người Việt Nam nói chung về biến này. Các kiểm định tham số dựa trên giá trị trung bình đã được dùng để chứng minh ảnh hưởng của các yếu tố xã hội – nhân khẩu học (tuổi, giới, địa dư, học vấn) lên chất lượng cuộc sống trước chấn thương.

Vì sự khác nhau giữa điểm của QoL trước chấn thương và điểm của QoL sau chấn thương có phân bố chuẩn nên kiểm định t ghép cặp được sử dụng để đánh giá sự thay đổi về chất lượng của cuộc sống trước và sau chấn thương. Điểm QoL được phân vào hai mức đủ (≥ 50); thấp (< 50) và kiểm định χ^2 McNemar được sử dụng để xác định sự thay đổi QoL.

Số ngày điều trị trong bệnh viện không phải là một phân bố chuẩn, ảnh hưởng của loại phương tiện sử dụng khi chấn thương như đi bộ, hay đi xe đến thời gian nằm viện đã được đánh giá bằng kiểm định phi tham số dựa trên các trung vị. Mối liên quan giữa số ngày nằm viện và QoL được kết luận thông qua hệ số tương quan Spearman.



Tỷ lệ chấn thương ở đầu/cột sống được so sánh với các số liệu năm 1997 và những người đi bộ được so sánh với những người sử dụng phương tiện giao thông khác trong nghiên cứu bằng kiểm định χ^2 . Kiểm định χ^2 cũng được sử dụng để xem liệu có phải điểm QoL khác nhau theo mức độ chấn thương khi lượng giá theo vị trí chấn thương nặng nhất.

Chú ý rằng phần phương pháp cần viết ngắn gọn và tránh sự lặp lại khi đã sử dụng tóm tắt tương tự và kiểm định nhiều lần trước đó. Những lý do tại sao lại dùng những kiểm định đã chọn cũng cần phải chỉ rõ (vì phân bố chuẩn hay đó là biến phân loại...).

KẾT QUẢ

Chất lượng cuộc sống

Điểm trung bình QoL trước chấn thương của các đối tượng trong nghiên cứu chấn thương do giao thông quốc gia là 58,0 (độ lệch chuẩn 0,2) cao hơn trung bình của toàn quốc (50,0) sự khác biệt này có ý nghĩa thống kê ($t_{1691} = 42,8, p < 0,001$).

Chưa thấy có mối liên quan giữa chất lượng cuộc sống trước chấn thương được lượng giá bằng điểm QoL với giới tính ($t_{1690} = 0,5, p = 0,486$) hay địa dư ($F_{7,1684} = 1,7, p = 0,116$) trong nghiên cứu. Tuy nhiên, ở nhóm trẻ dưới 6 tuổi và những người có trình độ học vấn trung học có chất lượng cuộc sống cao hơn (điểm trung bình tương ứng là 64,0 và 62,0) một cách có ý nghĩa thống kê so với những người có TĐHV cấp II hoặc dưới cấp II (điểm trung bình tương ứng là 58,0 và 55,0) ($F_{3,1688} = 52,1, p < 0,001$).

Đã có bằng chứng về sự giảm một cách có ý nghĩa của điểm trung bình chất lượng cuộc sống là 5,7 sau chấn thương so với trước chấn thương (từ 60,4 xuống 54,7) với khoảng tin cậy 95% của 5,4 đến 6,0 ($t_{1691} = 38,2, p < 0,001$). Không có đủ bằng chứng để kết luận tuổi (Pearson's $r = 0,24$) cũng như thời gian điều trị tại bệnh viện (Spearman's $r = 0,09$) có mối tương quan chặt chẽ với chất lượng cuộc sống sau chấn thương.

Thời gian điều trị tại bệnh viện.

Điểm trung vị số ngày điều trị tại bệnh viện của những đối tượng đi bộ là 5, ít hơn 2 ngày so với những đối tượng sử dụng các phương tiện khác ($Z = -1,96, n = 751, p = 0,05$). Thời gian điều trị tại bệnh viện cũng khác nhau một cách có ý nghĩa thống kê giữa những người sử dụng các loại phương tiện giao thông khác nhau khi bị tai nạn (Kruskal-Wallis test, $n = 660, p = 0,003$), điểm trung vị cao nhất (15 ngày) thuộc nhóm đối tượng sử dụng xe máy.

Chấn thương ở đầu/cột sống

Tỷ lệ các nạn nhân tai nạn giao thông bị chấn thương ở đầu/cột sống trong điều tra năm 2001 là 36%, điều này cho thấy không có bằng chứng về việc giảm tỷ lệ chấn thương ở đầu/cột sống so với điều tra năm 1997 là 37% ($\chi^2_1 = 0,2, p = 0,623$). Có 44% các nạn nhân đi bộ bị chấn ở đầu hoặc cột sống trong khi tỷ lệ này ở nhóm nạn nhân sử dụng các phương tiện khác chỉ là 37%, tuy nhiên sự khác biệt này không có ý nghĩa thống kê với mức $\alpha = 0,05$ ($\chi^2_1 = 3,3, p = 0,070$). Chưa có bằng chứng về việc vị trí thương tích ảnh hưởng đến chất lượng cuộc sống tương xứng sau chấn thương (> 50 điểm) ($\chi^2_2 = 2,1, p = 0,349$).

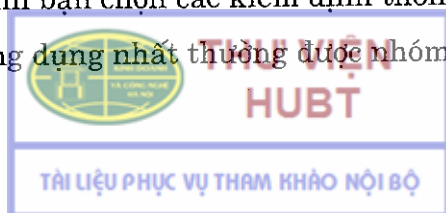
Những tiêu đề đã được sử dụng để phản ánh những phần khác nhau mà nhóm nghiên cứu quan tâm. Cần lưu ý rằng có một vài phần được phân tích riêng biệt đã được gộp lại và một vài phần phân tích lại được mô tả theo một trật tự khác đi để có thể mang lại một “câu chuyện” nhất quán hơn cho người đọc. Điều này là hoàn toàn thích hợp, bạn đang viết một báo cáo chứ không phải nhật ký! Trong bản báo cáo bạn nên thường xuyên nêu lên ý nghĩa của các kiểm định, ví dụ các trung bình khác nhau như thế nào, tỷ lệ ở nhóm nào là cao nhất cũng như các kiểm định thống kê đã dùng và mức ý nghĩa của chúng.

Bản báo cáo ví dụ này là một bản báo chỉ sử dụng các từ ngữ để mô tả các mối liên quan, tuy nhiên nếu sự khác nhau về chất lượng cuộc sống theo địa dư có ý nghĩa thống kê thì bạn có thể trình bày theo dạng bảng phân bố giá trị trung bình theo địa dư hoặc biểu đồ. Trong trường hợp này bảng nên đưa ra ngay trong phần kết quả mô tả của báo cáo và có thể được tham khảo trong phần viết về kết quả của các kiểm định thống kê. Cần phải cân nhắc cả phần mô tả và phần suy luận trong kết quả nghiên cứu phải bổ xung cho nhau và tránh sự chồng chéo không cần thiết.

4.7. Giả định

Tất cả các phần tóm tắt và các kiểm định thống kê đều có các giả định cần thiết và các giả định này phải đạt được nếu chúng ta muốn sử dụng các kết quả thống kê một cách chính xác. Sử dụng sai giá trị thống kê hoặc các kiểm định có thể dẫn đến những kết luận sai lầm. Trong mọi trường hợp, nếu bạn yêu cầu máy tính thực hiện một phân tích thống kê thì nó sẽ thực hiện ngay, kể cả khi kiểm định đó hoàn toàn không có giá trị. Là một người phân tích số liệu, bạn có trách nhiệm phải kiểm tra tất cả các giả định liên quan tới kiểm định thống kê và điều này đôi khi cần thiết bạn phải có những phân tích thêm. Phần tiếp theo đây sẽ cung cấp cho bạn cách phân tích cần thiết để kiểm tra cho hầu hết các giả định thông thường cần thiết cho các kiểm định thống kê được trình bày trong cuốn sách này. Bảng 1(tr.193) và 2(tr.194) sẽ cho bạn biết những giả định nào cần được cân nhắc khi bạn chọn các kiểm định thống kê.

Những giả định thông dụng nhất thường được nhóm như sau:



1. Với tất cả các kiểm định thống kê
 - Sự độc lập của các đơn vị quan sát.
2. Kiểm định thống kê liên quan với các biến phụ thuộc liên tục.
 - Phân bố chuẩn của biến phụ thuộc.
 - Tính đồng nhất của các biến trong nhóm so sánh ngang.
 - Không có bằng chứng về đa cộng tuyến.
3. Kiểm định thống kê liên quan đến biến phụ thuộc phân loại.
 - Giá trị kỳ vọng đủ lớn.

4.7.1. Sự độc lập của các đơn vị quan sát

Tất cả các kiểm định thống kê cơ bản trong chủ đề này yêu cầu giả định về *tính độc lập của các đơn vị quan sát* phải được thoả mãn. Điều đó có nghĩa là giá trị một biến phụ thuộc của một đối tượng nghiên cứu không chịu ảnh hưởng của giá trị của đối tượng khác. Với những thiết kế nghiên cứu dựa trên cách lấy mẫu ngẫu nhiên đơn giả định này thường được thoả mãn. Đôi khi, các đối tượng nghiên cứu có thể biết nhau (ví dụ trong trường hợp lấy mẫu kiểu snowball) hoặc các đối tượng nghiên cứu có thể cùng trong một gia đình, trường học, làng, cơ quan... dẫn đến các thành viên trong cùng cùng gia đình/trường học... có nhiều đặc điểm giống nhau khi đánh giá trong cụm hơn giữa những cá nhân từ những cụm khác nhau. Điều này dẫn đến những *sự phụ thuộc* chéo của một số đặc điểm. Những kiểm định thống kê bạn sẽ học trong chương trình này không thể đối phó với những mức độ phụ thuộc khác nhau giữa các đơn vị quan sát, có nhiều kiểm định phức tạp hơn có thể làm được điều này.

Để quyết định xem các giả định có thoả mãn không, bạn cần biết đến cách lấy mẫu của bộ số liệu đã được thu thập. Hãy chú ý những câu hỏi sau:

Có bằng chứng nào cho thấy rằng có sự co cụm của các cá nhân trong mẫu nghiên cứu, do đặc điểm tự nhiên (*gia đình, trường học, làng xóm*) hoặc chúng ta tạo ra (*lấy mẫu kiểu snowball*) không?

Nếu Có, thì giả định về tính độc lập của đơn vị quan sát có vẻ không thoả mãn và cần phải có cách tiếp cận khác – và trong trường hợp này bạn nên tham khảo ý kiến của các chuyên gia thống kê.

Chú ý rằng sự co cụm của các đối tượng quan sát trong cùng một đơn vị quan sát là chấp nhận được – điều này xảy ra trong các nghiên cứu đo lường lặp lại. Chỉ có các đơn vị quan sát là phải độc lập với nhau.

4.7.2. Phân bố chuẩn

Một trong những giả định cần phải thoả mãn khi phân tích các biến phụ thuộc dạng liên tục sử dụng giá trị trung bình là phân bố tần số của biến có phải là phân bố chuẩn không.

Câu hỏi liệu biến phụ thuộc có phân bố tần số theo phân bố chuẩn hay không có thể được chuyển thành “Chúng ta sẽ sử dụng trung bình hay trung vị để ước lượng giá trị thống kê?” Rất nhiều kiểm định thống kê yêu cầu giả định này phải được thoả mãn, để đơn giản việc tính toán, nhiều người sẽ sử dụng giá trị trung bình thay cho trung vị trong việc ước lượng trung bình. Giá trị trung bình chỉ có thể thay thế cho giá trị trung vị khi giả định về phân bố chuẩn được thoả mãn.

Có những phân tích thống kê sẽ giúp chúng ta xác định phân bố tần số của mẫu có là phân bố chuẩn hay không. Ví dụ, kiểm định Kolmogorov–Smirnov được biết đến là một kiểm định tính chuẩn. Tuy nhiên, việc kiểm tra một phân bố chuẩn hoàn hảo đôi khi không cần thiết vì chúng ta cũng chỉ cần kiểm tra phân bố đó có *xấp xỉ* phân bố chuẩn hay không mà thôi. Những kiểm định đưa ra trong bảng 1(tr.193) không yêu cầu phải có phân bố chuẩn hoàn hảo, chỉ cần xấp xỉ phân bố chuẩn. Vì thế thuật toán sau đây để đánh giá liệu giả định về phân bố chuẩn của biến phụ thuộc có được thoả mãn hay không sẽ được dùng để đánh giá tính chuẩn.

1. Tính toán những giá trị sau từ bộ số liệu: Giá trị trung bình, trung vị, độ lệch chuẩn, giá trị cực đại, cực tiểu, skewness, kurtosis, và biểu đồ cột liên tục. Dùng các phép tính thống kê để tính các giá trị này. Hãy xem ví dụ ở bài 3.

2. Nếu bạn có thể trả lời Có cho tất cả các câu hỏi sau, bạn đã có bộ số liệu xấp xỉ phân bố chuẩn.

- Giá trị trung bình có nằm trong 10% giá trị trung vị không?
- Giá trị trung bình $\pm 3sd$ có xấp xỉ giá trị cực đại và cực tiểu trong bộ số liệu không?
- Hệ số skewness có nằm trong $\pm 3sd$ không?
- Hệ số kurtosis có nằm trong $\pm 3sd$ không?
- Biểu đồ cột liên tục có xuất phát điểm thấp, cao nhất ở giữa sau đó thấp dần về phía xa (không cần thiết phải theo đúng hình chuông) không?
- Ngoài ra, nếu (và chỉ nếu) biến liên tục xuất phát từ giá trị 0 (đây không phải là giá trị phủ định), thì độ lệch chuẩn có ít hơn # giá trị trung bình không?




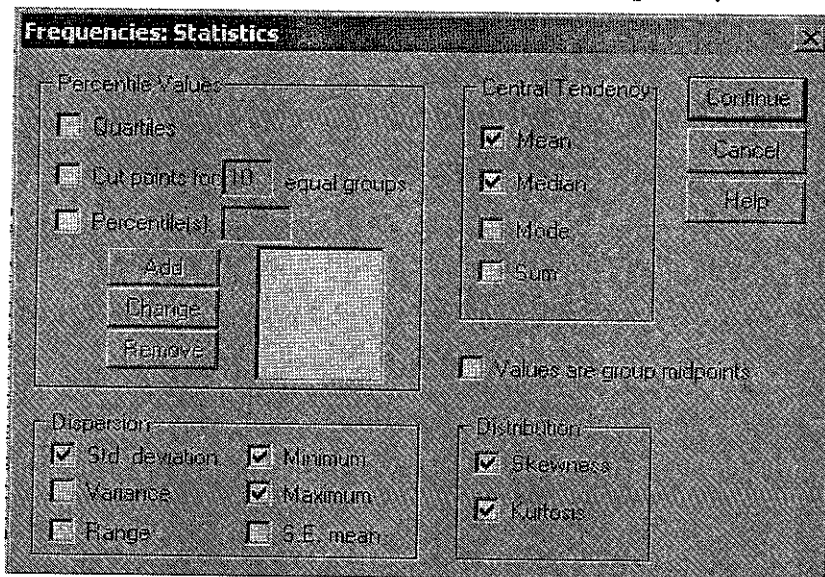
Nếu câu trả lời là có cho tất cả các tiêu chuẩn trên thì biến này xấp xỉ phân bố chuẩn.

Khi một biến liên tục tuân theo phân bố chuẩn, bạn có thể tính toán về giá trị trung tâm và sự phân tán của biến theo trung bình và độ lệch chuẩn. Nếu không phải là phân bố chuẩn bạn không thể sử dụng số trung bình nhưng có thể sử dụng giá trị trung vị và cực tiểu – cực đại hoặc những phân vị khác để mô tả sự phân tán.



SỬ DỤNG SPSS ĐỂ TÍNH TOÁN NHỮNG THÔNG TIN CẦN CHO ĐÁNH GIÁ PHÂN BỐ CHUẨN

1. Thực hiện theo các bước sau:
2. Từ thực đơn dọc chọn: **Analyse/Descriptive Statistics/Frequencies**
3. Từ danh sách biến, chọn biến `qol_bef` (Chất lượng chung của cuộc sống trước chấn thương) và chuyển vào hộp biến bằng cách nhấp chuột lên biểu tượng  tượng
4. Nhấp chuột lên nút **Statistics**, bạn sẽ thấy hộp thoại tương tự như hình dưới đây. Đánh dấu vào các hộp thống kê bạn cần – mean, median, std. dev., skewness, kurtosis, minimum, maximum – sau đó nhấp chuột vào **Continue**.



5. Nhấp chuột vào nút **Charts**, chọn **Histogram**, sau đó nhấp chuột vào **Continue**.

6. Nhấp chuột vào **OK**.

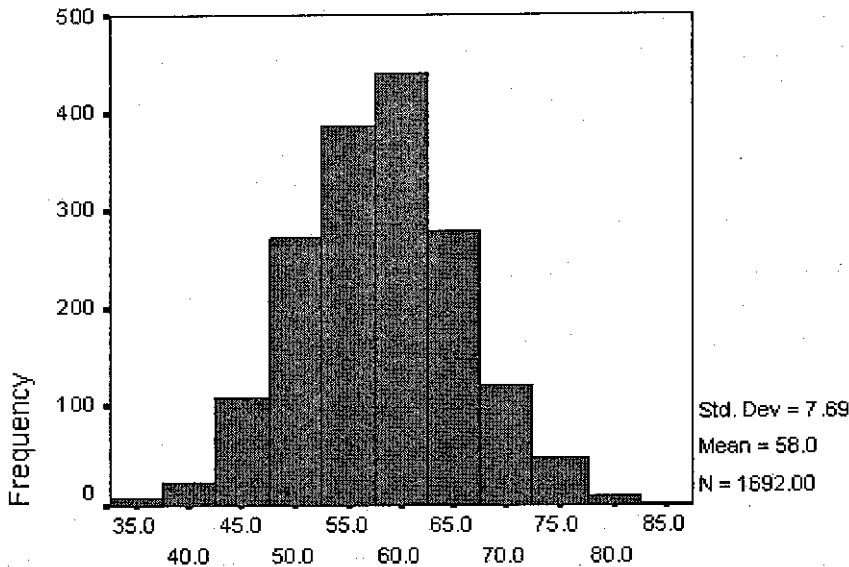
Kết quả sẽ xuất hiện riêng rẽ trong cửa sổ Window như trong phần kết quả dưới đây.

KẾT QUẢ

Statistics

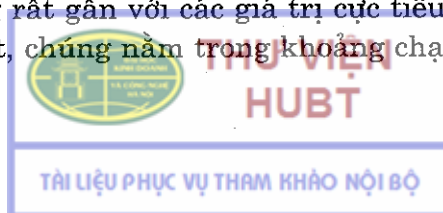
General quality of life before injury		
N	Valid	1692
	Missing	29
Mean		57.9941
Median		58.0000
Std. Deviation		7.68642
Skewness		.070
Std. Error of Skewness		.059
Kurtosis		.086
Std. Error of Kurtosis		.119
Minimum		34.00
Maximum		85.00

General quality of life before injury



General quality of life before injury

Từ kết quả này bạn có thể thấy rằng giá trị trung bình nằm trong 10% trung vị. Giá trị trung bình $\pm 3sd$ tương ứng là $(58,0 - 3 \times 7,69 = 35,0)$ và $(58,0 + 3 \times 7,69 = 81,0)$. Chúng rất gần với các giá trị cực tiểu 34 và cực đại 85. Hệ số skewness và kurtosis tốt, chúng nằm trong khoảng chạy từ $-3sd$ tới $+3sd$, biểu



đồ cột liên tục gần giống hình chuông. Vì thế giả định về phân bố chuẩn được chấp nhận. Giá trị trung bình và độ lệch chuẩn sẽ được sử dụng để kết luận về biến phụ thuộc này.

4.7.3. Tính đồng nhất của phương sai ở các nhóm so sánh

Cả kiểm định tham số và phi tham số đã mô tả trong quyển sách này chỉ có giá trị khi giả định về sự phân tán của giá trị biến phụ thuộc gần giống nhau trong các nhóm so sánh ngang. Vì thế giả định này cần được kiểm định khi bạn quan tâm đến các giả thuyết về các biến phụ thuộc liên tục và bao hàm ít nhất là hai nhóm so sánh.

Bạn sẽ so sánh phương sai của các nhóm so sánh, vì thế sẽ cần đưa ra độ lệch chuẩn, số cực tiểu và cực đại của biến phụ thuộc riêng biệt cho từng nhóm. Nếu sự phân tán thống kê giữa các nhóm so sánh gần như nhau thì đã có sự đồng nhất của phương sai. Một số kiểm định, như kiểm định mẫu độc lập t, đã được sửa đổi để đối phó với việc không thoả mãn giả định này— xem mục 4.7 để có thêm thông tin. Tuy nhiên, trong hầu hết các trường hợp khác, nếu bạn không thể chỉ ra được tính đồng nhất về phương sai giữa các nhóm so sánh bạn không thể sử dụng một cách có hiệu quả hầu hết các kiểm định tham số và phi tham số. Đôi khi sử dụng phương pháp đổi biến để chuyển đổi số liệu của các nhóm so sánh để có các giá trị phương sai đồng nhất và trong những trường hợp như vậy bạn nên tham khảo sách thống kê hoặc các chuyên gia thống kê.

Về hình thức, giả định này được kiểm định một cách tự động khi bạn sử dụng kiểm định Levene's trước khi dùng kiểm định t không ghép cặp, nhưng với những kiểm định thống kê khác mà không được kiểm định tự động bạn có thể sử dụng kiểm định F cho phương sai của hai nhóm. Kiểm định F dùng để kiểm định các giả thuyết mà tỷ số của các phương sai là 1 (không có sự khác nhau). Tham khảo thống kê Y tế Công cộng phần Thống kê cơ bản để có thêm thông tin về kiểm định F.

Không giống như giả định về phân bố chuẩn, giả định phân bố xấp xỉ chuẩn thì tính đồng nhất của phương sai giữa các nhóm so sánh rất quan trọng. Các kiểm định thực hiện rất nhạy cảm với việc không thoả mãn giả định này, bạn có thể có những kết luận sai lầm khi phân tích nếu bạn vi phạm giả định này.





SỬ DỤNG SPSS ĐÁNH GIÁ TÍNH ĐỒNG NHẤT CỦA PHƯƠNG SAI

Giả định bạn nghĩ đến giả thuyết rằng giá trị trung bình điểm QoL khác nhau theo loại phương tiện giao thông có liên quan đến chấn thương. Bạn có 5 nhóm phương tiện giao thông: xe ô tô, xe đạp, xe máy, người đi bộ, và loại phương tiện khác. Trước hết bạn cần đưa ra các giá trị trung bình và phương

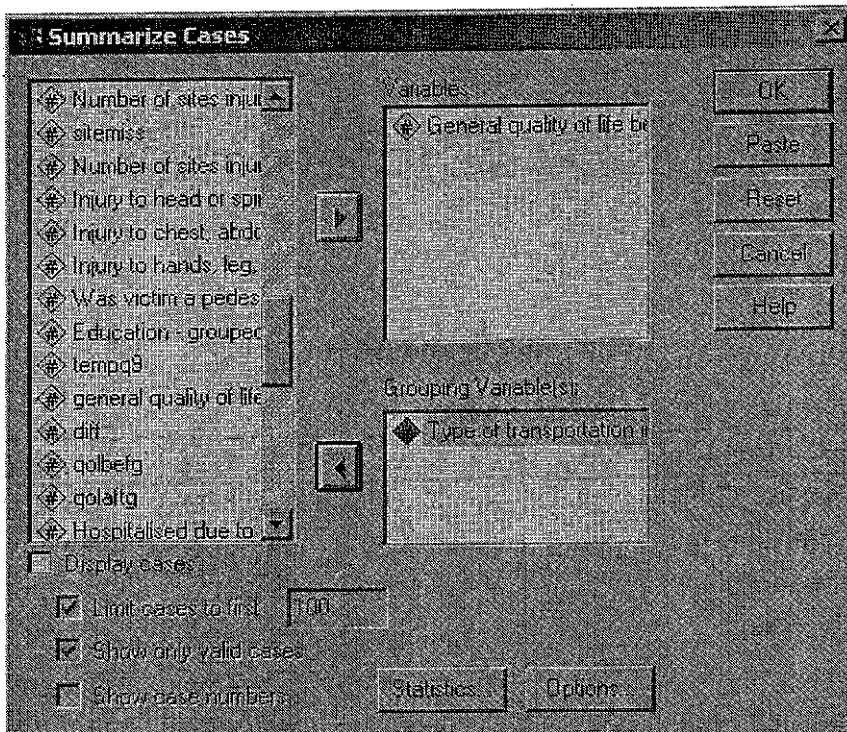
sai của điểm QoL trong từng nhóm phương tiện. Để làm được điều này thực hiện theo các bước sau:

1. Từ thực đơn dọc chọn: **Analyse/Reports/Case Summaries.**

2. Từ danh sách biến, nhấp chuột vào biến qol_bef (chất lượng chung của cuộc sống trước khi chấn thương) và chuyển vào hộp biến bằng cách nhấp chuột vào biểu tượng 

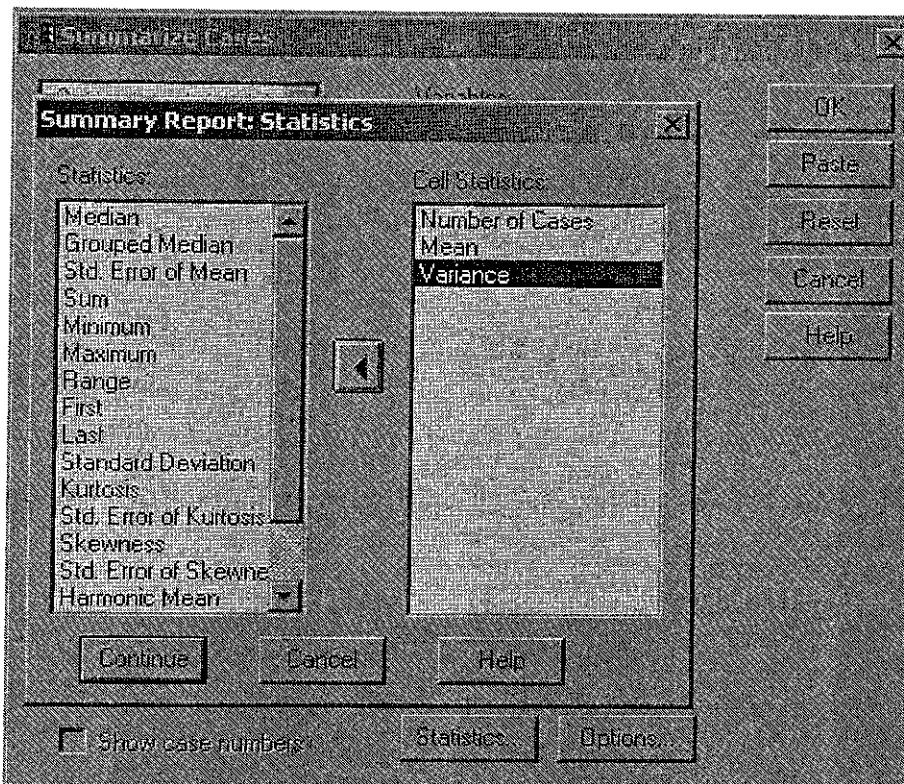
3. Từ danh sách biến, chọn nhóm biến, ví dụ chọn nhóm education, trantype, và chuyển vào hộp nhóm biến bằng cách nhấp chuột vào biểu tượng 

4. Kích chuột vào 'Display cases', SPSS đã đánh dấu mặc định, chuyển biến đã đánh dấu vào hộp



5. Nhấp chuột vào nút **Statistics**, chuyển lựa chọn **Mean** và **Variance** vào hộp **Cell Statistics** – kích vào **Continue/OK**. Màn hình sẽ tương tự như hình dưới đây:





Kết quả sẽ xuất hiện như sau

Case Summaries

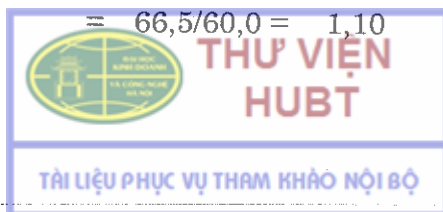
General quality of life before injury

Type of transportation in	N	Mean	Variance
motorised vehicle	68	59.0882	66.470
Bicycle	325	57.8769	62.880
Motorised bike	885	58.2169	60.014
Pedestrian	165	58.0909	56.912
Other	67	57.0448	51.498
Total	1510	58.1172	60.159

Chúng ta có thể thấy rằng phương sai của các loại là gần giống nhau. Tỷ số thống kê F được tính toán và trình bày rõ ràng hơn một chút. Không có tỷ suất nào có ý nghĩa thống kê với mức $p < 0,05$.

$$\text{Xe ô tô so sánh với xe đạp} = 66,5/62,9 = 1,06$$

$$\text{Xe máy} = 66,5/60,0 = 1,10$$



Người đi bộ	=	$66,5/56,9 = 1,17$
Loại khác	=	$66,5/51,5 = 1,29$
Xe đạp so sánh với xe máy	=	$62,9/60,0 = 1,05$
Người đi bộ	=	$62,9/56,9 = 1,11$
Loại khác	=	$62,9/51,5 = 1,22$
Xe máy so sánh với người đi bộ	=	$60,0/56,9 = 1,05$
Loại khác	=	$60,0/51,5 = 1,17$
Người đi bộ so sánh với loại khác	=	$56,9/51,5 = 1,10$

Tương tự như các kiểm định thống kê thông thường cho tính chuẩn, kiểm định F cũng chịu ảnh hưởng của cỡ mẫu, thậm chí đôi khi với tỷ suất tương đối nhỏ (dưới 1,5) cũng có thể có ý nghĩa thống kê. Chỉ cần giá trị thống kê F là 1,5 hoặc nhỏ hơn, bạn cũng có thể cho là giả định về tính đồng nhất của phương sai đã được thoả mãn.

4.7.4. Cộng tuyến

Khi sử dụng kiểm định thống kê để định lượng về độ mạnh của mối liên quan giữa hai biến, sẽ có lần chúng ta gặp phải hai biến có mối liên quan rất chặt chẽ đến nỗi từ giá trị của biến này chúng ta có thể biết được giá trị của biến kia. Ví dụ, số ngày mưa và số ngày nắng trong một tháng có thể dùng để dự đoán cho những tháng khác. Rất hiếm khi thấy nắng khi trời đang mưa, vì vậy tương quan giữa nắng và mưa là một tương quan nghịch rất mạnh. Khi các mối tương quan giữa hai biến mạnh đến mức 0,9 hoặc cao hơn nữa bạn nên xem xét xem liệu hai biến này trên thực tế có cùng lượng giá cho một hiện tượng không. Nếu chúng cùng lượng giá cho một hiện tượng thì chúng ta đã sai khi đưa chúng vào cùng một phân tích thống kê. Nếu chúng ta cứ phân tích chúng thì ta sẽ mắc phải một vấn đề là sự *cộng tuyến* trong việc phân tích.

Để kiểm tra sự cộng tuyến, chúng ta cần tính toán mối tương quan giữa hai biến mà ta nghi ngờ. Nếu hệ số tương quan là 0,9 hoặc cao hơn, không được tiếp tục dùng cả hai biến trong phân tích của bạn. Hãy chọn một trong hai biến đó, thường chọn biến dễ dàng lượng giá hơn và tiến hành phân tích chỉ với một biến này.

Lưu ý rằng với những nghiên cứu đo lường nhắc lại, bạn sẽ có những tương quan cao hơn giữa những đo lường nhắc lại của cùng một biến đã được một người đo rất nhiều lần. Những tương quan này tương đối được chấp nhận trong nghiên cứu đo lường nhắc lại.






SỬ DỤNG SPSS ĐÁNH GIÁ SỰ CỘNG TUYẾN

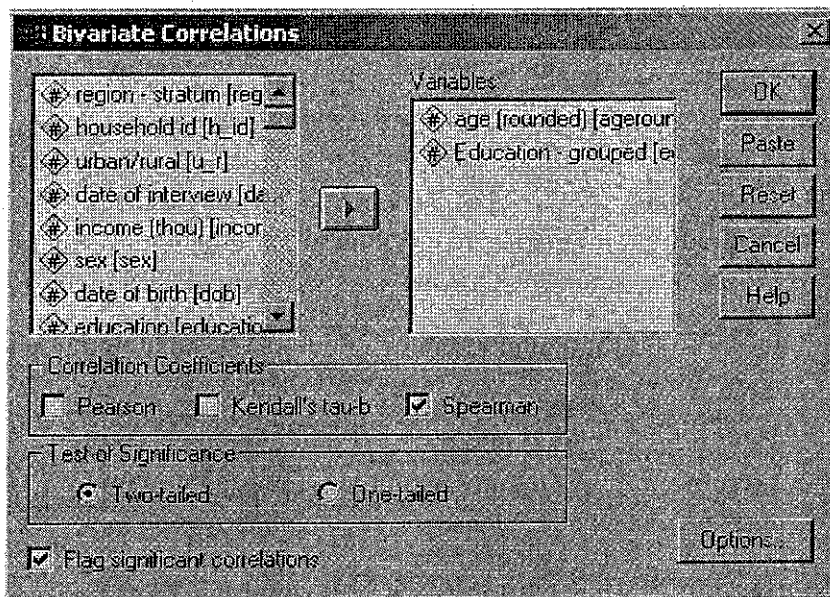
Giả sử rằng bạn đang nghi ngờ về khả năng có sự cộng tuyến giữa tuổi và trình độ học vấn (trình độ học vấn thường được tính dựa vào số năm học ở trường, tuy nhiên cũng có thể đo lường biến này bằng một đơn vị khác ví dụ là thời gian). Bạn cần phải chỉ ra mối tương quan của tuổi và trình độ học vấn. Tuổi là một biến liên tục và trình độ học vấn là biến thứ hạng với các giá trị như Trẻ nhỏ/mù chữ, Cấp 1, Cấp 2, Cấp 3. Hệ số tương quan Spearman's là hệ số tương quan thích hợp cho các dạng biến loại này (xem module 3). Để thực hiện được điều này trong chương trình SPSS bạn cần thực hiện theo những bước sau.

1. Từ thực đơn dọc chọn: **Analyse/Correlate/Bivariate.**

2. Từ danh sách biến, chọn biến **ageround** (tuổi) và **edgrp** (trình độ học vấn)

và chuyển vào hộp biến bằng cách nhấp chuột vào biểu tượng 

3. Nhấp chuột vào **Pearson's correlation** rồi chuyển bằng cách đánh dấu (SPSS chọn mặc định nhưng giá trị này không phù hợp), nhấp chuột lên **Spearman** để chọn dạng tương quan sẽ được tính.



4. Nhấp chuột vào **OK** để hoàn thành lệnh.

Kết quả sẽ xuất hiện riêng rẽ trong cửa sổ dưới đây :

Correlations

		age (rounded)	Education - grouped
Spearman's rho	age (rounded)	Correlation Coefficient	1.000
		Sig. (2-tailed)	.228**
		N	.000
Education - grouped	Education - grouped	Correlation Coefficient	1721
		Sig. (2-tailed)	.228**
		N	1702

** . Correlation is significant at the .01 level (2-tailed).

Trong trường hợp này, mặc dù có mối tương quan vừa phải giữa tuổi và trình độ học vấn nhưng hai biến này rõ ràng dùng để lượng giá những đặc điểm khác nhau. Trong khi những người trẻ nhất như trẻ em thường có trình độ học vấn thấp và số năm đến trường tăng tương ứng với tuổi ở một vài người trong quần thể, trường hợp cộng tuyến và mối tương quan chặt chỉ xảy ra khi các cơ hội học tập là giống nhau đối với tất cả mọi người trong quần thể. Trên thực tế lại không như vậy, một số người già không có trình độ học vấn cao như những người khác, vì thế không thể có được mối tương quan chặt chẽ.

Vì hệ số tương quan nhỏ hơn 0,9, sự cộng tuyến tính không phải là một vấn đề khi chúng ta đưa hai biến này vào trong cùng một phân tích.

4.7.5. Giá trị kỳ vọng đủ lớn

Khi xem xét mối liên quan giữa các biến phân loại, các bảng ngang biểu diễn kết quả của tần số và tỷ lệ và thường được phân tích với vài dạng của kiểm định χ^2 (xem phần 4.6.13). Các kiểm định χ^2 tính toán các giá trị kỳ vọng trong bảng với giả định là giả thuyết không (H_0) đúng, các giá trị kỳ vọng còn được sử dụng để ước lượng và là số liệu để tính toán giá trị thống kê. Mọi kiểm định χ^2 đều giả định là giá trị kỳ vọng của các ô phải từ 5 trở lên.

Lưu ý, giả định yêu cầu là các giá trị kỳ vọng lớn hơn 5 chứ không phải là các giá trị quan sát. Các giá trị quan sát của bạn có thể rất thấp thậm chí bằng 0.



Khi bạn sử dụng SPSS để tính toán kiểm định χ^2 , chương trình cũng có thể đưa ra giá trị kỳ vọng trong kết quả. Trừ khi bạn yêu cầu đưa ra các giá trị kỳ vọng, SPSS sẽ chỉ in ở phần dưới kết quả kiểm định χ^2 những thông tin cho bạn biết có bao nhiêu ô không thoả mãn điều kiện và giá trị kỳ vọng dưới 5,0. Dòng tin này xuất hiện trong hầu hết các phần kết quả của kiểm định χ^2 , và khi giả định không thoả mãn bạn không thể dùng kiểm định này. Nếu trường hợp đó xảy ra bạn cần phải xem xét việc gộp các giá trị phân loại liên kế của một biến hoặc cả hai, nhưng chỉ nên thực hiện khi việc gộp này có ý nghĩa với mối liên

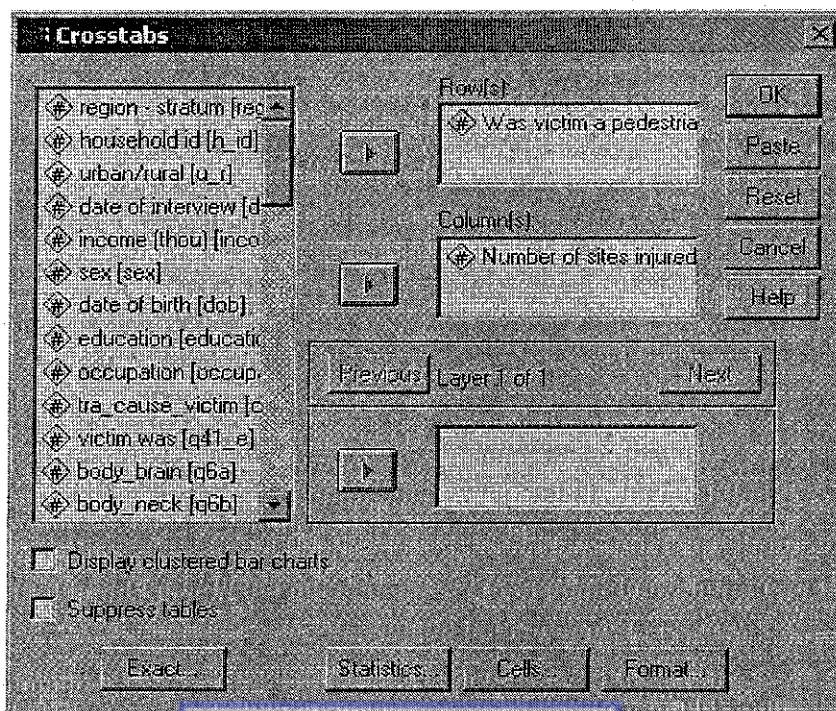
quan mà bạn đang tìm hiểu. Sau khi gộp các giá trị phân loại bạn thực hiện lại kiểm định χ^2 và xem xét xem giả định đã được thoả mãn chưa. Đôi khi bạn đã gộp rất nhiều các giá trị phân loại nhưng giả định về tần số kỳ vọng vẫn không thể thoả mãn. Nếu tần số kỳ vọng trong bảng nhỏ nhất là 2,0 hoặc lớn hơn, bạn có thể giải thích kết quả phân tích tuy nhiên nên đề cập đến việc này. Nếu tần số kỳ vọng dưới 2,0 thì bạn không thể sử dụng kết quả phân tích và chỉ có thể dùng các thống kê mô tả để phiên giải số liệu này.



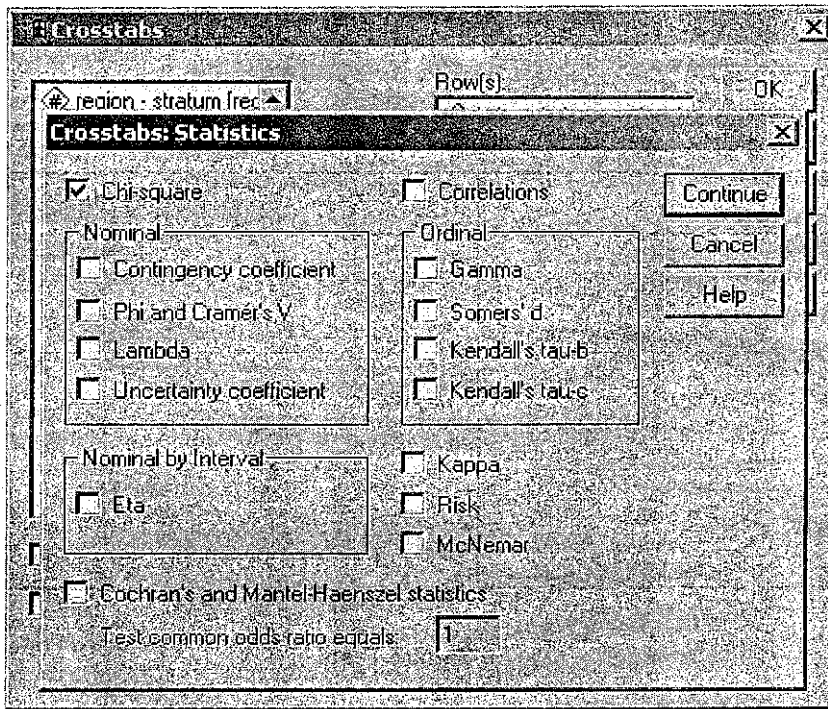
SỬ DỤNG SPSS ĐÁNH GIÁ ĐỘ LỚN CỦA GIÁ TRỊ KỲ VỌNG

Giả sử rằng bạn đang tìm hiểu mối liên quan giữa số lượng vị trí chấn thương và nạn nhân đi bộ khi bị chấn thương. Bạn thực hiện một kiểm định χ^2 để kiểm định giả thuyết này. Giả định về giá trị kỳ vọng đủ lớn được kiểm định sau khi bạn tiến hành kiểm định χ^2 và thông tin này được trình bày trong phần kết quả. Bạn thực hiện theo các bước sau

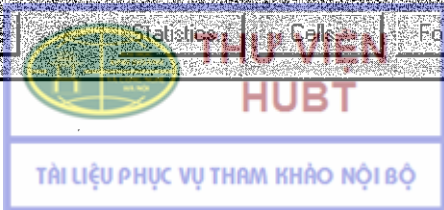
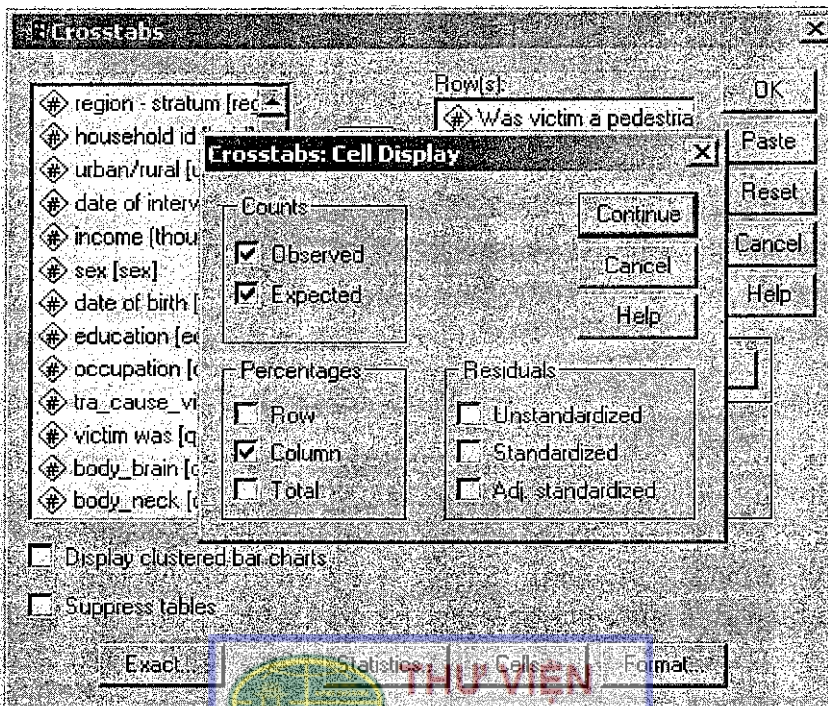
1. Từ thực đơn dọc chọn: Analyze/Descriptive Statistics/Crosstabs.
2. Từ danh sách biến, chọn biến pedestrn (người đi bộ) và chuyển vào hộp Row(s) bằng cách nhấp chuột vào biểu tượng .
3. Từ danh sách biến, chọn biến sitesg (số lượng vị trí chấn thương) và chuyển vào hộp Column(s) bằng cách nhấp chuột vào biểu tượng .



4. Nhấp chuột lên nút Statistics và chọn Chi Square .



5. Nhấp chuột lên Continue trở về màn hình chính, kích vào nút Cells. SPSS đã sẵn sàng chọn Observed. Bạn phải chọn Expected, và Column để đưa ra bảng giá trị kỳ vọng và tỷ lệ theo cột.



6. Nhấp chuột lên Continue để trở về màn hình chính.

7. Cuối cùng nhấp chuột lên nút OK để kết thúc lệnh.

Kết quả sẽ xuất hiện như sau:

Was victim a pedestrian? * Number of sites injured - grouped Crosstabulation

			Number of sites injured - grouped				Total
			0	1	2	3+	
Was victim a pedestrian?	No	Count	43	992	253	69	1357
		Expected Count	40.0	1002.8	250.0	64.1	1357.0
		% within Number of sites injured - grouped	95.6%	88.0%	90.0%	95.8%	89.0%
	Yes	Count	2	135	28	3	168
		Expected Count	5.0	124.2	31.0	7.9	168.0
		% within Number of sites injured - grouped	4.4%	12.0%	10.0%	4.2%	11.0%
Total	Count	45	1127	281	72	1525	
	Expected Count	45.0	1127.0	281.0	72.0	1525.0	
	% within Number of sites injured - grouped	100.0%	100.0%	100.0%	100.0%	100.0%	

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6.811 ^a	3	.078
Likelihood Ratio	8.262	3	.041
Linear-by-Linear Association	1.909	1	.167
N of Valid Cases	1525		

a. 1 cells (12.5%) have expected count less than 5. The minimum expected count is 4.96.

Tần số quan sát nhỏ nhất là 2. Tần số kỳ vọng nhỏ nhất là 4,96 theo như thông tin cuối cùng ở trên. Điều này cho thấy giả định về tần số kỳ vọng từ 5,0 trở lên gần như đã thoả mãn. Kiểm định χ^2 có thể có giá trị sử dụng trong trường hợp này.

4.7.6. Kết luận

Như bạn đã thấy, việc lựa chọn phân tích thống kê phụ thuộc vào rất nhiều các hiểu biết chi tiết về câu hỏi nghiên cứu, thiết kế nghiên cứu. Một khi bạn có một ý tưởng rõ ràng về những gì sẽ cần trong bản báo cáo cuối cùng bạn mới có thể xây dựng kế hoạch cho phân tích số liệu. Cũng như việc chuẩn bị cho nhiều việc, dành nhiều thời gian cho việc lập kế hoạch phân tích từ khi mới bắt đầu nghiên cứu sẽ tiết kiệm rất nhiều thời gian cho bạn khi phân tích sau này.



Bài số

5

NHIỄU VÀ KHỔNG CHẾ NHIỄU

Mục tiêu

1. Trình bày được các khái niệm về nhiễu và việc sai lệch kết quả nghiên cứu khi có nhiễu.
2. Trình bày những kiến thức và báo cáo được về sự khác nhau giữa phân tích thô và phân tích hiệu chỉnh.
3. Trình bày được các nguyên tắc về sự hiệu chỉnh một giá trị thống kê.

5.1. Giới thiệu

Hầu hết các nghiên cứu sức khỏe liên quan đến các lượng giá quan sát trên con người trong môi trường sống tự do, có ít tác động vào đó. Điều này dẫn đến một khả năng lớn về sự khác biệt giữa các cá nhân che lấp bất kỳ sự khác biệt nào do can thiệp của chúng ta hoặc sự khác biệt thật sự giữa các nhóm. Các phân tích thống kê cần phải tính đến những khác biệt này càng nhiều càng tốt trước khi tìm kiếm sự khác biệt thật sự mà chúng ta quan tâm. Những khác biệt phức tạp này được biết đến như là các tác động nhiễu.

5.2. Nhiễu

5.2.1. Định nghĩa nhiễu

Nhiễu trong nghiên cứu y tế công cộng còn có nghĩa là có giải thích khác xen vào trong kết quả nghiên cứu. Điều này có nghĩa là mối liên quan giữa biến phụ thuộc và biến độc lập có thể là thật nhưng cũng có thể là có một vài mối liên quan khác là nguyên nhân của kết quả này. Sự có mặt của nhiễu dẫn đến những kết luận nghiên cứu sai, vì vậy nhiễu cần phải được khống chế.

Ví dụ, các kết quả có thể chỉ ra trình độ học vấn có liên quan đến chất lượng cuộc sống sau chấn thương, nhưng trên thực tế có thể là tuổi mới thực sự ảnh hưởng đến chất lượng cuộc sống. Vì có mối liên quan giữa tuổi với trình độ học



THƯ VIỆN
HUBT

TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

vấn, nên khi có mối liên quan giữa số tuổi của một người với điểm QoL thì đường như cũng có mối liên quan giữa trình độ học vấn và QoL. Mối liên quan của trình độ học vấn và QoL được xem như bị nhiễu bởi biến tuổi.

Những định nghĩa dịch tễ của nhiễu là có sự mất cân bằng của các đặc tính trong các nhóm so sánh, ví dụ một nhóm có biểu hiện đặc tính đặc thù của một người được so sánh với người khác trong nhóm hoặc nhóm khác.

Định nghĩa thống kê của nhiễu là một biến thứ ba có tương quan với cả biến phụ thuộc và các biến độc lập mà chúng là mối quan tâm hàng đầu trong kiểm định giả thuyết này.

Trong nghiên cứu y tế công cộng, tuổi là một biến nhiễu cổ điển, tất cả các điều kiện sức khỏe đều chịu ảnh hưởng của tuổi tác, và rất nhiều biến độc lập cũng cùng thay đổi với tuổi.

5.2.2. Khống chế nhiễu khi thiết kế nghiên cứu

Các tác động của nhiễu có thể được khống chế trong khâu thiết kế nghiên cứu và phân tích số liệu. Thiết kế nghiên cứu cẩn thận sẽ cân nhắc cả cách tiếp cận cũng như khả năng thực hiện. Phân bố ngẫu nhiên và ghép cặp là hai cách thường dùng để khống chế nhiễu một cách tối đa trong thiết kế nghiên cứu.

5.2.2.1. Phân bố ngẫu nhiên

Như đã trình bày trong phần 1, thiết kế nghiên cứu thực nghiệm sử dụng khống chế hầu hết khi trả lời một câu hỏi nghiên cứu. Với sự phân bố ngẫu nhiên, các nhà nghiên cứu làm tối đa cơ hội để phân bố các đặc tính vào đều trong các nhóm so sánh. Chỉ có duy nhất một điều khác biệt giữa các nhóm là có một hay nhiều can thiệp được tiến hành cho nhóm đó. Bất kỳ sự khác biệt nào ở kết quả trong phân tích cũng có thể là do can thiệp chứ không phải tác động của nhiễu.

5.2.2.2. Ghép cặp

Có nhiều câu hỏi nghiên cứu trên thực tế không thể trả lời được khi sử dụng thiết kế nghiên cứu thực nghiệm, và cần phải sử dụng đến các thiết kế nghiên cứu quan sát. Trong hầu hết các đề tài nghiên cứu, danh sách các biến nhiễu dự kiến thường được đưa ra từ trước khi làm nghiên cứu; qua việc xem xét các nghiên cứu đã tiến hành trước đây trên thực địa hoặc những kiến thức chuyên môn của nhóm nghiên cứu. Nếu chúng ta biết rằng một hoặc hai biến có ảnh hưởng đến mối liên quan được nghiên cứu, thì có thể thiết kế nghiên cứu sử dụng phương pháp ghép cặp với biến nhiễu.

Ghép cặp bao gồm sự bắt buộc một nhóm so sánh có những đặc tính tương tự như nhóm khác. Ví dụ, nếu biến tuổi được cho là biến nhiều với mối liên quan giữa trình độ học vấn và chất lượng cuộc sống, thì chúng ta thiết kế một nghiên cứu sao cho phân bố tuổi trong tất cả các nhóm trình độ học vấn là như nhau. Nếu chúng ta lấy mẫu theo phương pháp ngẫu nhiên đơn từ quần thể, chúng ta có thể chắc rằng số người trẻ tuổi ở nhóm trình độ học vấn dưới THCS sẽ nhiều hơn nhóm trình độ học vấn từ THCS trở lên. Nếu chúng ta ghép cặp các nhóm so sánh, chúng ta nên nhóm theo phân bố tuổi của nhóm có trình độ học vấn từ THCS trở lên với nhóm dưới THCS với mẫu là những người trẻ hơn trước. Điều này được biết đến như ghép cặp theo tần suất, khi các đặc tính của nhóm bắt buộc là phải tương tự nhau.

Ghép cặp cũng có thể được thực hiện ở mức độ cá thể, thường thấy trong thiết kế nghiên cứu bệnh – chứng. Ví dụ, bệnh ung thư vòm họng có khuynh hướng gặp ở quần thể những người lớn tuổi và thường gặp ở nam giới. Để đảm bảo rằng sự phân bố giới tính và tuổi là như nhau, ghép cặp cá thể có thể chọn ngẫu nhiên một chứng cho một trường hợp ung thư từ tập hợp quần thể chứng với tuổi và giới tính tương tự như trường hợp bệnh nghiên cứu.

Bằng ghép cặp, tần suất hoặc cá thể, các ảnh hưởng của biến nhiều cũng được khống chế, vì mối liên quan của chúng với biến phụ thuộc đã được loại bỏ (cả nhóm chứng và nhóm bệnh có phân bố như nhau về các biến nhiều nên không có nhiều như các khái niệm ở trên).

5.2.3. Khống chế nhiều khi phân tích số liệu

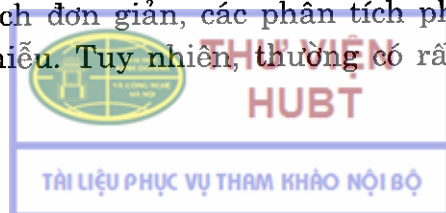
Trong phân tích, có hai cách để khống chế các tác động của các biến nhiều: phân tích phân tầng và mô hình hồi quy đa biến.

5.2.3.1. Phân tích phân tầng

Phân tích phân tầng trong giai đoạn phân tích sẽ cho ta kết quả tương đương với ghép cặp trong thiết kế nghiên cứu. Phân tầng có nghĩa tiến hành phân tích hai biến phụ thuộc và biến độc lập trong theo các phân nhóm của biến nhiều. Bằng cách phân tầng và xem xét trong phân nhóm của biến nhiều bạn làm cho mọi đối tượng trong từng phân nhóm có mối liên quan với biến nhiều tương tự như nhau, vì thế biến này không còn là biến nhiều nữa theo như định nghĩa trong phần 5.2.1.

5.2.3.2. Mô hình hồi quy đa biến

Với những phân tích đơn giản, các phân tích phân tầng thích hợp hơn trong việc thăm dò nhiều. Tuy nhiên, thường có rất nhiều biến nhiều cần



phải khổng chế, phân tích phân tầng rất nhanh chóng trở nên đơn điệu – khi cứ phải lặp lại sự phân tích hai biến trong rất, rất nhiều phân nhóm. Một cách nâng cao hơn để khổng chế đa nhiễu là sử dụng mô hình hồi quy đa biến. Mô hình này dạng mở rộng của hồi quy tuyến tính đơn giản, mô hình này sử dụng nhiều hơn một biến độc lập để giải thích sự thay đổi trong một biến phụ thuộc. Mặc dù các mô hình đa biến đều có thể làm được điều tương tự (điều chỉnh đa nhiễu), trên thực tế có rất nhiều dạng khác nhau của mô hình đa biến. Giống như sự lựa chọn của kiểm định thống kê cơ bản trong bài 4, lựa chọn mô hình đa biến dựa trên độ đo của biến phụ thuộc. Trong phần này sẽ chỉ đề cập đến một dạng của mô hình đa biến (hồi quy đa tuyến tính), nhưng nguyên lý có thể được khái quát hoá cho tất cả các mô hình khác. Giống như một phân tích phương sai cơ bản, chúng phân chia sự biến thiên của biến phụ thuộc thành các biến thiên thành phần: ảnh hưởng của nhóm (biến độc lập), biến nhiễu 1, biến nhiễu 2, v.v. Kiểm soát các ảnh hưởng của nhiễu cho phép có một lượng giá “tinh” hơn về ảnh hưởng của các biến độc lập, và sự ảnh hưởng này được coi là các **ước lượng** có hiệu chỉnh. Ước lượng mà chúng ta hay dùng trước đây vẫn thường được gọi là **ước lượng** hay ước lượng không hiệu chỉnh. “Ước lượng” là thuật ngữ thông thường cho một giá trị thống kê được rút ra từ phân tích như – trung bình, tỷ lệ, tỷ suất chênh, tương quan...

5.2.4. Bài tập ví dụ

Hãy xem xét mối liên quan trong ví dụ về, mối liên quan giữa QoL sau chấn thương và trình độ học vấn được nghi ngờ là bị nhiễu bởi biến tuổi. Để đơn giản, chúng ta hãy xem xét trình độ học vấn theo hai nhóm (“dưới THCS” và “bằng hoặc trên THCS”), và nhóm tuổi theo 3 nhóm (0–14, 15–49, và 50–100 tuổi). Câu hỏi nghiên cứu là “Có mối liên quan gì giữa QoL sau chấn thương và trình độ học vấn?”. Câu hỏi này được đưa ra dưới dạng giả thuyết khoa học

H₀: Điểm trung bình QoL sau chấn thương là như nhau không liên quan đến trình độ học vấn.

5.2.4.1. Phân tích không hiệu chỉnh

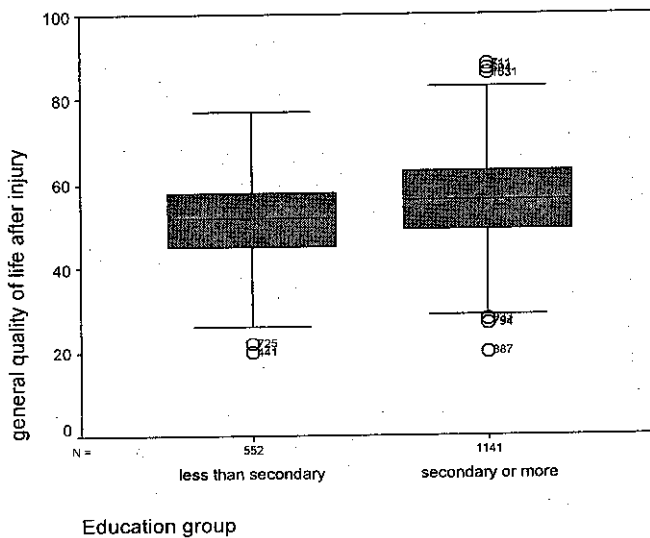
Phân tích đầu tiên được trình bày là kiểm định t so sánh hai số trung bình về điểm QoL trong hai nhóm trình độ học vấn.

Group Statistics

Education group	N	Mean	Std. Deviation	Std. Error Mean
general quality of life after injun less than secondar	552	51.6214	9.52386	.40536
secondary or more	1141	56.1402	9.89327	.29288

Independent Samples Test

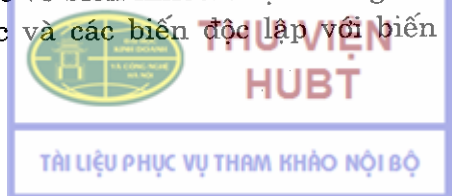
	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
general quality of life after inju	.729	.393	-8.917	1691	.000	-4.5189	.50677	-5.51281	-3.52490
			-9.036	127.888	.000	-4.5189	.50010	-5.50008	-3.53762



Từ các kết quả mô tả và phân tích, cho thấy sự khác biệt có ý nghĩa thống kê về điểm trung bình QoL theo trình độ học vấn ($p < 0,001$). Điểm trung bình QoL ở nhóm có trình độ học vấn từ THCS trở lên cao hơn nhóm dưới THCS là 4,5 điểm (95% khoảng tin cậy 3,5 đến 5,5).

5.2.4.2. Tuổi có phải là biến nhiễu không?

Định nghĩa thống kê về biến nhiễu được chứng minh là có các mối liên quan giữa cả biến phụ thuộc và các biến độc lập với biến nhiễu tiềm tàng. Vì thế



chúng ta cần tìm xem biến tuổi (nhiều tiềm tàng) có liên quan với (i) điểm QoL sau chấn thương (biến phụ thuộc) và (ii) trình độ học vấn (biến độc lập).

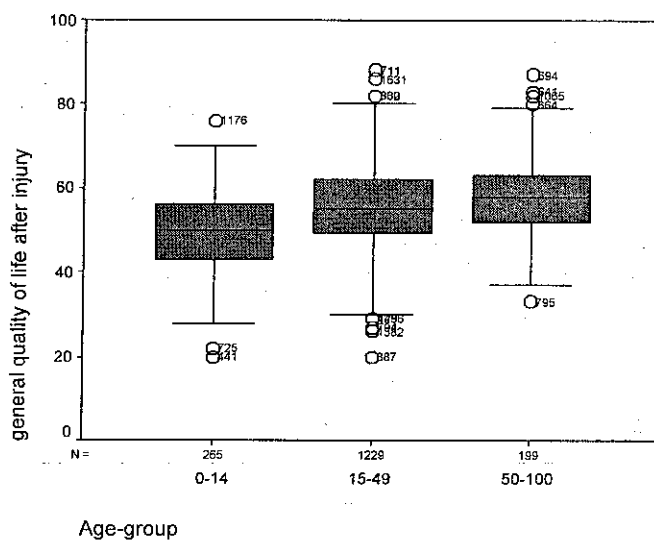
Điểm QoL sau chấn thương và nhóm tuổi.

QoL là biến liên tục và tuổi là biến phân nhóm trong ví dụ này. Vì thế tính trung bình điểm QoL trong từng nhóm tuổi.

Report

general quality of life after injury

Age-group	Mean	N	Std. Deviation
0-14	49.6981	265	9.56751
15-49	55.2278	1229	9.88559
50-100	57.8191	199	9.01051
Total	54.6669	1693	9.99864



Có khuynh hướng tăng điểm QoL khi tuổi tăng. Vì thế tuổi và QoL là có liên quan với nhau.

Lưu ý: Không thích hợp để sử dụng ý nghĩa thống kê để đưa ra kết luận về một biến có là nhiều hay không. Sử dụng ý nghĩa ngữ cảnh trong trường hợp này.

Trình độ học vấn và nhóm tuổi, cả hai đều là biến phân loại, vì thế bảng ngẫu nhiên được sử dụng:

Education group * Age-group Crosstabulation

		Age-group			Total
		0-14	15-49	50-100	
Education group	less than secondary	Count 201	271	85	557
	% within Age-group	75.6%	21.9%	42.7%	32.7%
secondary or more	Count	65	966	114	1145
	% within Age-group	24.4%	78.1%	57.3%	67.3%
Total	Count	266	1237	199	1702
	% within Age-group	100.0%	100.0%	100.0%	100.0%

Tỷ lệ các đối tượng có trình độ học vấn từ THCS trở lên ở nhóm tuổi 15-49 tuổi cao hơn cơ bản so với hai nhóm tuổi còn lại, và đúng như mong đợi, tỷ lệ này ở nhóm tuổi dưới 15 thấp hơn nhóm tuổi 50-100.

Từ (i) và (ii), chúng ta đã xác minh được biến tuổi có liên quan đến cả biến phụ thuộc và biến độc lập, vì thế biến tuổi cần phải được xem như là biến nhiều của mối liên quan này và phải được điều chỉnh.

5.2.4.3. Thăm dò nhiều của biến tuổi sử dụng phân tích phân tầng

Phân tích phân tầng sẽ giúp thăm dò biến tuổi gây nhiều đến mối liên quan giữa điểm QoL và trình độ học vấn như thế nào. Như một phân tích lặp lại mối liên quan giữa hai biến trong phân nhóm của biến nhiều. Loại phân tích này là không thể thực hiện với các biến nhiều là dạng biến liên tục (vì thế phải dùng nhóm tuổi như trong ví dụ).

Report

general quality of life after injury

Age-group	Education group	Mean	N	Std. Deviation
0-14	less than secondary	49.0249	201	9.74651
	secondary or more	51.8125	64	8.71939
	Total	49.6981	265	9.56751
15-49	less than secondary	52.4023	266	9.26506
	secondary or more	56.0083	963	9.91376
	Total	55.2278	1229	9.88559
50-100	less than secondary	55.3176	85	8.13021
	secondary or more	59.6842	114	9.21457
	Total	57.8191	199	9.01051
Total	less than secondary	51.6214	552	9.52386
	secondary or more	56.1402	1141	9.89327
	Total	54.6669	1693	9.99864



Nhìn vào tất cả các dòng tổng trong bảng trên, chúng ta thấy mối liên quan không thích ứng của phần 5.3.4.1; sự khác nhau về điểm trung bình QoL giữa nhóm trình độ học vấn “dưới THCS” và nhóm “từ trên THCS” là 4,5 điểm. Hãy xem sự khác nhau tương tự ở các nhóm tuổi khác nhau; sự khác nhau là 2,8 điểm cho nhóm tuổi 0 đến 14, 3,6 điểm cho nhóm tuổi 15 đến 49, và 4,3 điểm cho nhóm tuổi 50 đến 100.

Đây là tất cả về sự sắp xếp tương tự, sự khác biệt là cao hoặc thấp hơn 3 trong từng trường hợp, và với phần lớn mẫu đều thấp hơn ước lượng không thích ứng là 4,5. Vì thế ước lượng không hiệu chỉnh có vẻ như là đã ước lượng quá với sự khác nhau thực sự. Bằng phân tích phân tầng, chúng ta không chế được ảnh hưởng của biến tuổi (ít nhất là với mức độ nhóm tuổi). Ước lượng có giá trị hơn về sự khác nhau của trung bình giữa hai nhóm trình độ học vấn khác nhau được dựa trên trọng số về sự khác nhau trung bình trong từng tầng theo tuổi, trọng số theo tỷ lệ trong từng tầng tuổi. Điều này mang lại điểm trung bình là 3,6, nó cơ bản là nhỏ hơn giá trị không hiệu chỉnh 4,5.

Bạn có thể thấy rằng nếu có nhiều biến nhiều đáng quan tâm, thì chúng ta sẽ nhanh chóng trở nên tẻ nhạt khi thực hiện các phân tích phân tầng riêng biệt và số liệu trong các bảng nhiều hàng nhiều cột (phức tạp) cũng trở lên rất thừa thớt. Phân tích phân tầng rất quan trọng khi suy xét hai hoặc ba biến nhiều trong bất kỳ nghiên cứu nào. Các phân tích cho bạn thấy điều đang xảy ra với mối liên quan mà không phải lo lắng đến những phức tạp về toán học và các giả định kết hợp là mô hình đa biến. Tuy nhiên, hầu hết các phân tích của nghiên cứu quan sát sẽ cần kết hợp vài mô hình đa biến để tính toán một cách đầy đủ những ảnh hưởng của biến nhiều.

5.2.4.4. Không chế nhiều bằng mô hình đa biến

‘Không chế’ một biến thường có nghĩa là các mối liên quan được xem xét dưới giả định là biến thứ ba (biến bị không chế) có giá trị như nhau cho mọi người trong tất cả các nhóm so sánh. Ở trên chúng ta đã không chế biến tuổi bằng cách phân tích những người nhóm tuổi 0–14 riêng rẽ với những người nhóm tuổi 15–49 và cũng riêng rẽ với những người nhóm tuổi 49–100. Có rất nhiều phương pháp dùng để không chế hoặc điều chỉnh ảnh hưởng của biến thứ ba đến mối liên quan giữa hai biến khác. Có một số tính toán dễ dàng có thể được thực hiện đơn giản bằng máy tính bấm tay, tuy nhiên một số khác lại yêu cầu tính toán trên máy vi tính.

Cách thông dụng và mạnh về thống kê trong không chế ảnh hưởng của một hoặc nhiều biến nhiều là hợp với các mô hình đa biến (như đã biết, không đúng trong hầu hết các trường hợp, như các mô hình đa biến trong một vài cuốn sách và tài liệu nghiên cứu). Trong hầu hết các phiên giải nghiên cứu về sức khỏe,

thường có nhiều hơn một biến nhiễu tiềm năng, và không phải tất cả các biến có hai hoặc ba phân loại. Ví dụ, trong một nghiên cứu quan tâm đến mối liên quan giữa trình trạng suy dinh dưỡng của trẻ và các thực hành trên đồng ruộng ở khu vực sông Mê Kông, nơi mà có giả thuyết cho rằng các thực hành trong tuổi tiêu đã lọc bỏ một số chất dinh dưỡng quan trọng từ đất và cây trồng ở đây thiếu sắt, đặc biệt cần cho sự phát triển của trẻ. Tuy nhiên, gần như sự phát triển của trẻ cũng như những thực hành trên đồng ruộng ở vùng này đều tương quan với thu nhập, trình độ học vấn của cha mẹ, sự tiếp cận với cơ sở y tế và rất nhiều biến khác nữa. Hãy hình dung số lượng các tính toán liên quan trong phân tích phân tầng cho tất cả các biến nhiễu tiềm tàng này! Mô hình đa biến sẽ không chế tất cả các biến nhiễu tiềm tàng này chỉ trong một lần.

Các dạng của các mô hình đa biến là đa dạng, phụ thuộc, tương tự như các kiểm định thống kê đơn giản bạn đã học, cho một dạng biến kết quả. Tuy nhiên, tất cả các mô hình đều có chung một mục đích – cung cấp các ước lượng không bị nhiễu của mối liên quan giữa hai biến. Mô hình dựa trên các kết quả sẽ được trình bày theo một cách giống nhau như các kết quả không hiệu chỉnh, hai biến đơn giản. Ví dụ, nếu bạn trích dẫn các giá trị trung bình trong kế hoạch phân tích hai biến thì mô hình bạn chọn cho mô hình đa biến nên đưa ra các giá trị trung bình hiệu chỉnh.

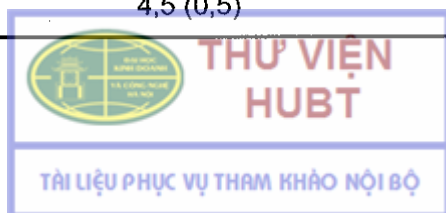
Mô hình đa biến điều chỉnh nhiễu:

- Cho các giá trị trung bình hiệu chỉnh sử dụng các mô hình hồi quy đa tuyến tính, nếu các mở rộng của hồi quy tuyến tính trong phần 4.5.15.
- Cho các tỷ lệ/tỷ suất hiệu chỉnh dùng các mô hình hồi quy Logistic.

Với ví dụ ở trên, mô hình hồi quy đa tuyến tính đã được cân nhắc là phù hợp với các phân bố rời rạc của trình độ học vấn và tuổi với điểm QoL sau chấn thương. Vì các phân bố của các biến này là rời rạc nên có thể có giá trị trung bình điểm QoL trong từng phân nhóm trình độ học vấn, không chế theo tuổi.

Bảng 5.1. Điểm trung bình QoL sau chấn thương trong các nhóm trình độ học vấn, hiệu chỉnh theo tuổi

	Trung bình hiệu chỉnh (se)	Trung bình chưa hiệu chỉnh (se)
Dưới THCS	51,6 (0,4)	52,1 (0,4)
Từ THCS trở lên	56,1 (0,3)	55,9 (0,3)
Khác biệt	4,5 (0,5)	3,8 (0,5)



Chú ý sự khác biệt hiệu chỉnh cho các biến nhiễu là do ước lượng mà có. Giá trị trung bình khác nhau giảm từ 4,5 xuống 3,8. Nhìn chung, khi các kết quả hiệu chỉnh là khác nhau căn bản với các ước lượng không hiệu chỉnh (cho là 10% hoặc hơn), đó là các kết quả hiệu chỉnh gần hơn với mối liên quan thật sự. Ở đây sự khác nhau giữa giá trị trung bình hiệu chỉnh và không hiệu chỉnh là hơn 10% sự khác biệt, chúng ta nên hiểu ở đây có nhiều thực sự của biến tuổi. Vì thế, các phân tích với nhiễu không được khống chế sẽ tạo ra các ước lượng bị sai số. Khi các kết quả hiệu chỉnh tương tự như các kết quả thô thì không có vấn đề gì cần nói về các kết quả. Đôi khi khống chế các biến mà chúng có thể trở thành các biến nhiễu cũng không làm thay đổi các ước lượng không hiệu chỉnh (không có nhiễu).

Nhiều có thể tác động theo hai hướng:

- Đôi khi những sự khác nhau đã hiệu chỉnh lại thấp hơn chưa hiệu chỉnh (nhiều dương tính).
- Đôi khi những sự khác nhau đã hiệu chỉnh lại cao hơn chưa hiệu chỉnh (nhiều âm tính).

Bài số

6

TÍNH CỖ MẪU

Mục tiêu

1. Trình bày được cách tiếp cận thông thường để tính toán cỡ mẫu cần thiết cho một câu hỏi nghiên cứu.
2. Trình bày được những khai niệm thiết kế nghiên cứu cần thiết cho việc tính toán cỡ mẫu.
3. Viết được cách tính cỡ mẫu cho một đề cương hoặc báo cáo nghiên cứu.

6.1. Các yếu tố ảnh hưởng đến tính tin cậy trong các kết quả

6.1.1. Ý nghĩa thống kê và ý nghĩa thực tế

Ngoài việc tăng tối đa khả năng câu hỏi nghiên cứu sẽ được trả lời một cách không chính xác, một mục đích khác của việc thiết kế nghiên cứu là đảm bảo không lãng phí nguồn lực. Điều này có thể dựa vào số lượng mẫu trong nghiên cứu, sẽ rất lãng phí khi mẫu nghiên cứu là 1000 trong khi chỉ cần 100 là đủ và tương tự nghiên cứu có thể là không có giá trị nếu mẫu nghiên cứu chỉ là 10 người trong khi phải cần tới 100 người. Làm thế nào để có thể xác định được mẫu là bao nhiêu thì đủ? Thực ra, nghĩa của từ “đủ” là gì?

Mẫu nghiên cứu là đủ khi kết quả nghiên cứu là đáng tin cậy và được chấp nhận. Điều đó có nghĩa là nếu kết quả đưa ra có thể chỉ được sự khác nhau giữa hai nhóm thì chúng ta cần chắc chắn rằng việc giải thích này không giống như bị ảnh hưởng đến của những dao động khi lượng giá. Chúng ta mong muốn tránh được kết quả dương tính hoặc âm tính giả.

Thông thường, một kết quả sẽ không được đề cập trong báo cáo trừ khi kết quả so sánh “có ý nghĩa thống kê, $p < 0,05$ ”. Chúng ta thường làm nghiên cứu và tìm ra một sự khác nhau mà lại không có ý nghĩa hoặc điều trái ngược, chúng ta có thể tìm thấy một kết quả khác nhau có ý nghĩa thống kê nhưng không có ý nghĩa trong thực tế. Điều này có thể là nhỏ, sự khác nhau không có ý nghĩa thống kê của một can thiệp trên thực tế có thể có ý nghĩa y tế công cộng lớn nếu sự thay đổi nhỏ đó tác động đến toàn bộ quần thể. Ngược lại, một can thiệp có

thể dẫn đến một sự khác biệt lớn có ý nghĩa thống kê cao nhưng lại không thích hợp nếu nó chỉ có thể áp dụng cho một số ít người trong quần thể.

Giải thích về phân tích thống kê nên dựa chủ yếu trên ý nghĩa thực tế, không phải trên ý nghĩa thống kê. Điều này thường không được trình bày rõ trong các sách thống kê. Giá trị p có thể chỉ được sử dụng để tuyên bố về “ý nghĩa” trong khi thiết kế nghiên cứu bao gồm cả tính toán cỡ mẫu cho một giả thuyết được kiểm định. Trong phạm vi cỡ mẫu, một kết luận âm tính giả là do cỡ mẫu quá nhỏ, và một kết luận dương tính giả là do cỡ mẫu quá lớn (nên nhớ rằng còn có các nguyên nhân khác trong thiết kế, như sai số chọn và sai số đo lường cũng có thể là nguyên nhân dẫn đến kết luận nghiên cứu sai).

Quá trình tính toán cỡ mẫu cho một công trình nghiên cứu là để chắc chắn rằng chúng ta đạt được cỡ mẫu đủ lớn dựa trên đó chúng ta đưa ra kết luận về vấn đề sức khỏe.

Điều đầu tiên của bất kỳ một tính toán cỡ mẫu nào cũng là những trình bày về kết quả của nghiên cứu này có tầm quan trọng như thế nào đối với sức khỏe cộng đồng. Ví dụ, chúng ta hy vọng rằng chiến dịch tuyên truyền tăng cường sức khỏe sẽ tăng nhận thức về chiến lược phòng chống HIV/AIDS lên 20% hoặc hơn nữa. Nếu sự cải thiện chỉ là 12%, nó không đủ hiệu quả để tiến hành chiến dịch này một cách thường xuyên, sự thay đổi này không đủ lớn với những chi phí nguồn lực đã bỏ ra. Vì thế mục đích của chúng ta là phải thiết kế nghiên cứu với cỡ mẫu đủ lớn để có thể phát hiện sự cải thiện tăng 20% hoặc hơn nữa có ý nghĩa thống kê. Chúng ta sẽ không quan tâm những sự khác biệt nhỏ hơn 20% mà không có ý nghĩa thống kê.

6.1.2. Biến thiên trong đo lường

Sự biến thiên trong đo lường là kết quả của biến thiên cá thể và biến thiên từng nhóm nếu chúng ta so sánh các nhóm. Nếu sự biến thiên trong đo lường giữa các cá thể lớn thì chúng ta sẽ khó để phát hiện những sự khác biệt nhỏ.

Đặt hai khái niệm trên (sự khác biệt nhỏ nhất mà có thể tìm ra và sự biến thiên) vào với ví dụ về điểm chất lượng cuộc sống trong bộ số liệu nghiên cứu của chúng ta, ta có thể có giả thuyết rằng điểm chất lượng cuộc sống sau chấn thương là khác nhau giữa nhóm nạn nhân có thu nhập thấp và nhóm có thu nhập không thấp. Sự khác biệt trung bình là 5 hoặc cao hơn là mối quan tâm của các nhà nghiên cứu. Hãy đơn giản hoá ví dụ này, giả định rằng lấy mẫu gồm 10 người trong mỗi nhóm thu nhập (thấp và không thấp). Sự khác biệt là 5 rất dễ dàng nhận ra trong việc so sánh tập hợp thứ nhất và tập hợp thứ 2 của các kết quả:

Điểm chất lượng cuộc sống của 10 người trong hai nhóm thu nhập

- Thu nhập thấp: 31 31 32 33 34 35 36 38 40 41
- Thu nhập không thấp 36 36 38 39 39 40 42 44 45 46

Điểm chất lượng cuộc sống không thay đổi nhiều (khoảng là 10 điểm) và sự khác biệt trung bình 5 là thuyết phục – mặc dù hai nhóm có sự chông chéo vẫn có sự thay đổi rõ ràng, điểm chất lượng cuộc sống ở nhóm thu nhập không thấp cao hơn nhóm thu nhập thấp.

So sánh với một ví dụ mà sự khác biệt được lượng giá cao hơn.

Điểm chất lượng cuộc sống của 10 người trong hai nhóm thu nhập:

- Thu nhập thấp: 31 32 34 38 40 41 46 48 49 51
- Thu nhập không thấp: 35 36 41 43 46 46 48 52 53 55

Trong trường hợp này, sự khác biệt trung bình vẫn là 5, nhưng kết quả không được chấp nhận vì sự khác nhau quá lớn của các cá thể trong nhóm (khoảng là 20) và sự chông chéo giữa hai nhóm cũng lớn.

Xem xét ví dụ về nhóm 20 người.

Điểm chất lượng cuộc sống của 20 người trong hai nhóm thu nhập:

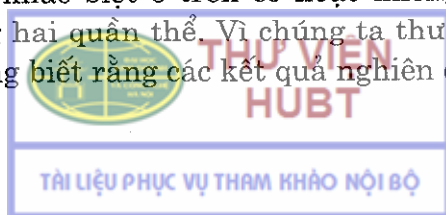
- Thu nhập thấp: 31 32 33 33 34 35 36 38 41 42 43 44 46 46 47 48 48 49 50 51
- Thu nhập không thấp: 39 41 42 42 43 44 45 46 47 49 49 50 51 52 53 53 54 55 55 56

Với cỡ mẫu lớn hơn, chúng ta tin tưởng hơn rằng sự khác nhau giữa hai nhóm là thật sự; sự “thay đổi” giá trị chất lượng cuộc sống cao hơn là nhất quán hơn và được dựa trên số lượng nhiều hơn.

Với thống kê mô tả, chúng ta thường muốn đưa ra khoảng tin cậy cho độ chính xác của các ước lượng thống kê (trung bình, tỷ lệ, tỷ suất chênh...). Sự khác biệt càng lớn, cỡ mẫu sẽ càng phải lớn để đưa ra số trung bình như nhau với độ chính xác như nhau.

6.1.3. Sai lầm loại I và sai lầm loại II

Kể cả khi có sự khác nhau thật sự tồn tại trong hai mẫu đối tượng, chúng ta vẫn có thêm một vấn đề nữa ảnh hưởng đến độ tin cậy của các kết quả. Điều này liên quan đến việc các đối tượng trong mẫu mà chúng ta chọn có đại diện cho toàn bộ quần thể hay không. Nếu hai nhóm đại diện được cho các quần thể của chúng thì sự khác biệt ở trên có thể sẽ được lặp lại (nếu chúng ta chọn lặp lại nhiều lần thì sự khác biệt đó vẫn sẽ tồn tại). Nếu hai nhóm này không đại diện cho quần thể, sự khác biệt ở trên có hoặc không thể phản ánh đúng sự khác biệt thật sự trong hai quần thể. Vì chúng ta thường không lặp lại nghiên cứu, nên chúng ta không biết rằng các kết quả nghiên cứu của chúng ta có phản



ảnh chính xác sự thật hay là có mắc phải sai lầm. Có hai loại sai lầm khi phiên giải kết quả chúng ta có thể mắc; sai lầm loại I và sai lầm loại II. Khái niệm về sai lầm loại I và sai lầm loại II là tương đương với khái niệm kết quả dương tính giả và âm tính giả trong kiểm định lâm sàng.

Mẫu	Sự thực (quần thể)	
	Các nhóm khác nhau	Các nhóm như nhau
Các nhóm khác nhau	✓	✗ (dương tính giả)
Các nhóm như nhau	✗ (âm tính giả)	✓

Nếu chúng ta đưa kết luận có sự khác biệt trong mẫu nghiên cứu và quần thể mà mẫu đại diện cũng có sự khác biệt này, chúng ta không có sai lầm trong kết luận.

Nếu chúng ta nói rằng không có sự khác biệt trong mẫu nghiên cứu và quần thể mà mẫu đại diện cũng không có sự khác biệt, chúng ta cũng không có sai lầm trong kết luận.

Nếu chúng ta đưa ra sự khác biệt trong mẫu nghiên cứu nhưng thực tế quần thể mà mẫu đại diện lại không có sự khác biệt này, chúng ta đã phạm phải *sai lầm loại I*.

Nếu chúng ta nói rằng không có sự khác biệt trong mẫu nghiên cứu, nhưng trên thực tế quần thể mà mẫu đại diện lại có sự khác biệt, chúng ta đã phạm phải *sai lầm loại II*.

Sai lầm loại I thường được cho là nghiêm trọng hơn sai lầm loại II. Vì khi chúng ta nói rằng có sự khác biệt nhưng trên thực tế kết luận của chúng ta được đưa ra từ một mẫu "tồi" còn tệ hại hơn là đưa ra kết luận là không có sự khác biệt. Kết luận này và bảng phía trên có thể được đưa vào phần kiểm định giả thuyết:

H_0 : Thời gian hoàn thành trung bình giữa hai nhóm là như nhau.

H_1 : Thời gian hoàn thành trung bình giữa hai nhóm là khác nhau.

Mẫu	Quần thể	
	H_1	H_0
H_1	✓	✗ (Sai lầm loại I)
H_0	✗ (Sai lầm loại II)	✓

Lực của kiểm định là phần bù của sai lầm loại II. Nếu sai lầm loại II là 10%, lực kiểm định là 90%.

6.1.4. Các mối quan hệ tương hỗ

Một cỡ mẫu được coi là đủ có thể nhỏ hơn 10 người hoặc lớn hơn 100000 người. Cỡ mẫu phụ thuộc vào mục đích của phân tích thống kê là mô tả hay suy luận, nếu là suy luận thì giả thuyết thống kê được kiểm định, và sự khác nhau tối thiểu có thể nhận thấy là mối quan tâm của các nhà nghiên cứu, đo lường của biến phụ thuộc và phương sai.

Cỡ mẫu tăng khi:

- Độ chính xác yêu cầu của ước lượng tăng.
- Sự khác nhau tối thiểu có thể nhận thấy giảm.
- Độ lệch chuẩn tăng.
- Sai lầm loại I hoặc sai lầm loại II giảm.
- Độ lượng giá trở nên tinh vi hơn (từ liên tục trở thành nhị thức).

6.2. Những điều kiện cần thiết để tính cỡ mẫu

Những gợi ý trên đây giúp cho bạn nắm được các cấu phần cần thiết cho việc tính cỡ mẫu. Tuy nhiên chúng ta cũng có những công thức giúp bạn ước lượng cỡ mẫu cần thiết để cân bằng cả 3 yếu tố có thể tác động đến độ tin cậy của kết quả nghiên cứu. Trước khi bạn sử dụng một trong những công thức đó bạn cần phải xác định rõ nghiên cứu của bạn thuộc loại *nghiên cứu mô tả* hay *nghiên cứu phân tích*.

Nếu mục đích của bạn là *nghiên cứu mô tả*, bạn sẽ quan tâm đến những kết quả có sự chính xác ở mức độ cao, và mục đích của tính toán cỡ mẫu để chắc chắn rằng cỡ mẫu của bạn đủ để đưa ra những kết quả này. Tất cả những gì bạn cần làm là đưa ra mức độ chính xác cho các tính toán.

Tuy nhiên, nếu mục đích của bạn là kiểm định giả thuyết (*thống kê suy luận*) bạn sẽ phải xác định các yếu tố sau trước khi tính toán cỡ mẫu:

- Có ý tưởng nào về đo lường biến thiên (độ lệch chuẩn) của biến phụ thuộc không.
- Có khả năng tìm ra sự khác biệt nhỏ nhất giữa hai nhóm so sánh.
- Nêu rõ mức độ của sai lầm loại I và sai lầm loại II mà bạn chấp nhận trong nghiên cứu của mình.

Hầu hết các nhà nghiên cứu cho phép 5% sai lầm loại I (bạn cũng cần chỉ rõ bạn muốn kiểm định một phía hay hai phía), và 10 % sai lầm loại II.



Điều này có nghĩa rằng bạn đã có sẵn một vài ý tưởng về kết quả nghiên cứu ngay cả khi bạn chưa thực hiện nghiên cứu! Thông thường trong những tài liệu có sẵn hoặc tiến hành nghiên cứu thí điểm sẽ cung cấp cho bạn một số thông tin về độ lệch chuẩn. Không ai có thể đưa ra được chuẩn để xác định sự khác biệt nhỏ nhất mà bạn muốn tìm ra vì điều này phụ thuộc vào từng nghiên cứu.

Kiểm định thống kê bạn chọn để phân tích phụ thuộc vào kiểu của biến phụ thuộc và hình thức so sánh (bài 4), việc lựa chọn công thức tính cỡ mẫu cũng phụ thuộc các yếu tố trên. Trên thực tế, công thức tính cỡ mẫu dựa trên sự biến đổi toán học của công thức kiểm định thống kê mà chúng đã được sử dụng trong sách này. Có những công thức khác nhau dựa trên dạng biến phụ thuộc khác nhau (biến liên tục biểu thị bằng số trung bình, biến phân loại biểu thị bằng tỷ lệ). Cũng có những công thức tính cỡ mẫu khác nhau phụ thuộc vào loại thiết kế nghiên cứu (ví dụ: nghiên cứu cắt ngang, nghiên cứu đo lường nhắc lại, so sánh trước sau...).

Bài này đề cập đến việc tính toán cỡ mẫu sử dụng chương trình phần mềm cho các thiết kế nghiên cứu thực nghiệm và nghiên cứu sử dụng phương pháp thu thập mẫu ngẫu nhiên đơn. Loại thiết kế này rất hiếm khi được áp dụng trong nghiên cứu y tế công cộng nơi mà các nghiên cứu quan sát và lấy mẫu cụm thường hay được sử dụng hơn. Bất kỳ một thiết kế nghiên cứu phức tạp nào (ví dụ: mẫu cụm, mẫu phân tầng, thiếu tính ngẫu nhiên) nên ước lượng tăng cỡ mẫu, cách điều chỉnh được đề cập tóm tắt trong phần 6.3.2.

6.3. Tính cỡ mẫu

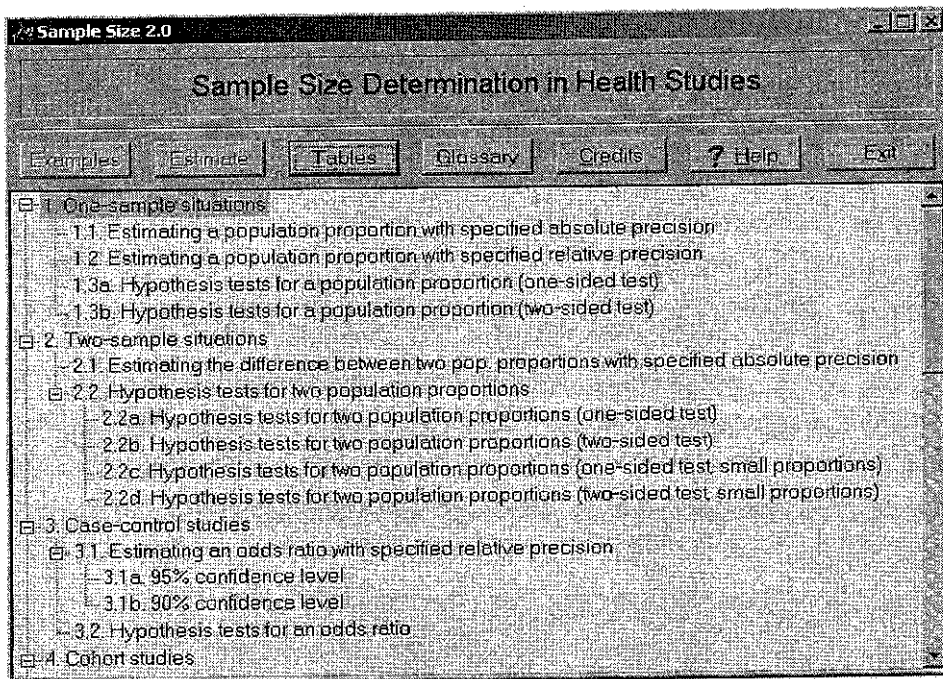
Cỡ mẫu cho rất nhiều loại thiết kế nghiên cứu và kiểm định giả thuyết có thể được tính trong phần mềm SSize, đây là một phần mềm miễn phí do Tổ chức y tế Thế giới phát triển. Phần mềm này cho phép tính toán cỡ mẫu cho rất nhiều loại thiết kế nghiên cứu và các loại giả thuyết khác nhau. Để cung cấp cho bạn cách sử dụng phần mềm này chúng tôi đưa ra 5 ví dụ dưới đây, hai ví dụ dựa trên thống kê mô tả và 3 ví dụ dựa trên thống kê suy luận; một ví dụ cho giả thuyết về nghiên cứu đo lường lặp lại (trung bình ghép cặp), một so sánh giữa hai trung bình của hai nhóm khác nhau, và một là so sánh hai tỷ lệ. Những ví dụ này phản ánh những câu hỏi từ bộ số liệu về chấn thương, nhưng hãy giả định rằng bộ số liệu này chưa được thu thập.

Cùng với việc cho bạn kết quả cỡ mẫu tính toán được, phần mềm cũng cung cấp cho bạn các công thức tính toán đã được dùng.

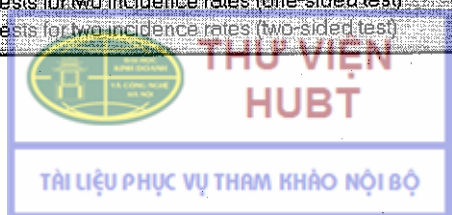
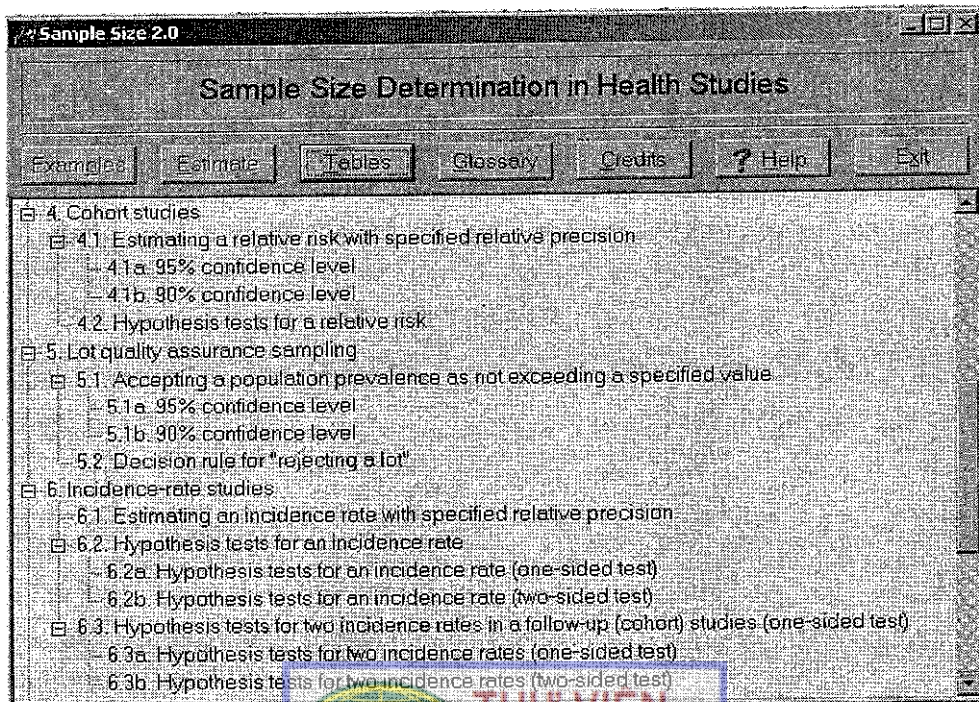
6.3.1. Những ví dụ về sử dụng SSize

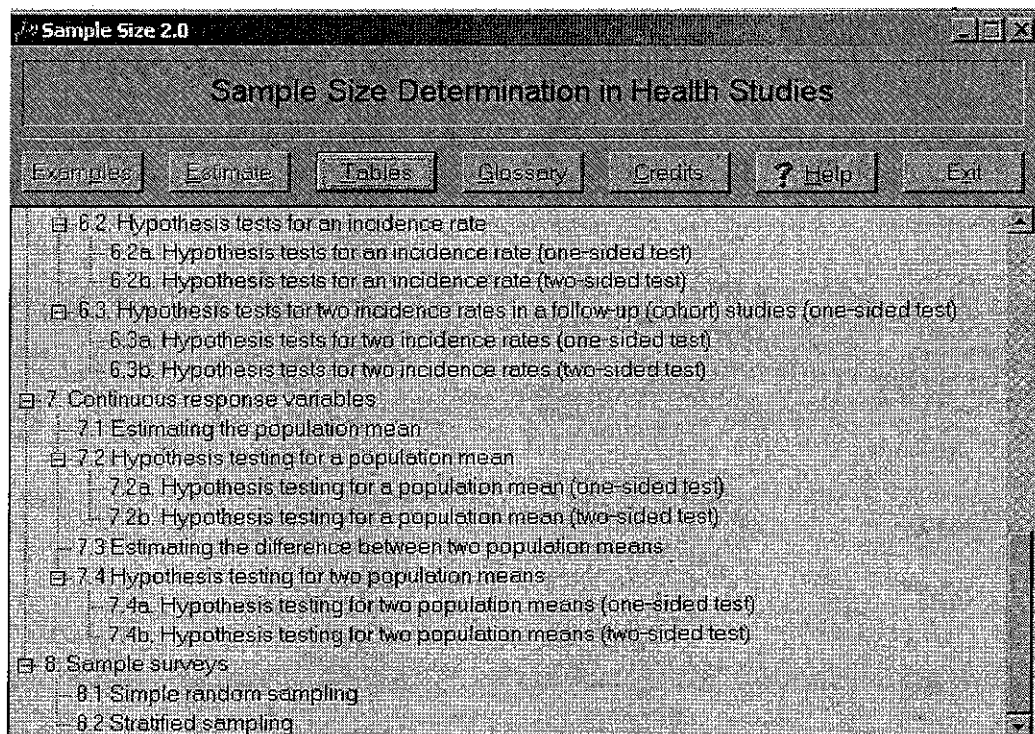
Vào SSize và bạn sẽ thấy một màn hình như sau:





Như bạn thấy, có rất nhiều khả năng tính toán để chọn lựa, và chỉ có một phần của các lựa chọn này được trình bày tiếp trong các màn hình tiếp theo:





Để có thể chọn được đúng cách tính toán phù hợp bạn cần phải hiểu rõ về kế hoạch phân tích của bạn. Nếu bạn không thể viết được những mô tả chi tiết hoặc những giả thuyết khoa học cho câu hỏi nghiên cứu của bạn thì bạn sẽ không thể lựa chọn được cách tính toán phù hợp giữa rất nhiều lựa chọn trên. Hãy xem bài 3 về phát triển kế hoạch phân tích.

Với các ví dụ, hãy giả định rằng nghiên cứu NTIS (National Transportation Injury Survey) vẫn ở giai đoạn thiết kế, và các nhà nghiên cứu muốn đảm bảo tính tin cậy của thống kê mô tả, và cỡ mẫu phù hợp cho kiểm định ba giả thuyết nghiên cứu.

6.3.1.1. Độ tin cậy của một ước lượng trung bình

Giả thuyết của câu hỏi số 4 ở bài 3 quan tâm đến ước lượng về chất lượng cuộc sống trước chấn thương. Thống kê mô tả đã được đưa ra cho biến này bao gồm cả khoảng tin cậy để phản ánh độ tin cậy của ước lượng trung bình.

H_0 : Điểm trung bình QoL trước chấn thương tương tự như quần thể chung, là 50.

Dựa trên các tài liệu có sẵn về công cụ lượng giá chất lượng cuộc sống cho thấy ở một quần thể đặc trưng có điểm QoL trung bình là 50 và độ lệch chuẩn là 10. Các nhà nghiên cứu đã kiểm định và xác định rằng điểm chất lượng cuộc

sống là phân bố chuẩn (xem mục 4.7). Các nhà nghiên cứu muốn đảm bảo rằng ước lượng điểm trung bình của chất lượng cuộc sống từ nghiên cứu NTIS có độ tin cậy là ± 5 điểm (điều này có nghĩa là điểm trung bình của quần thể nằm trong khoảng tin cậy 95% không lớn hơn ± 5).

Theo mong muốn của họ ước lượng trung bình quần thể là 50 với độ lệch chuẩn là 10 và độ tin cậy là ± 5 .

Để tính cỡ mẫu cần thiết ta làm như sau:

1. Từ thực đơn trên màn hình 6.3, chọn 7.1, nhấp chuột lên Estimate.

Màn hình tiếp theo sẽ hiển thị với các hộp trống trừ hộp $1-\alpha$. Độ tin cậy của 95% ở đây là quy ước và giả định. Bạn có thể thay đổi nếu cần thiết.

Perform Estimation

7.1. Estimating a population mean

Please select the desired unknown:

- Confidence level (%)
- Absolute precision required
- Relative precision
- Population mean
- Population standard deviation
- Population variance
- Sample size

Please enter the remaining values:

$1 - \alpha$ 95

d

ϵ

μ

σ

σ^2

n

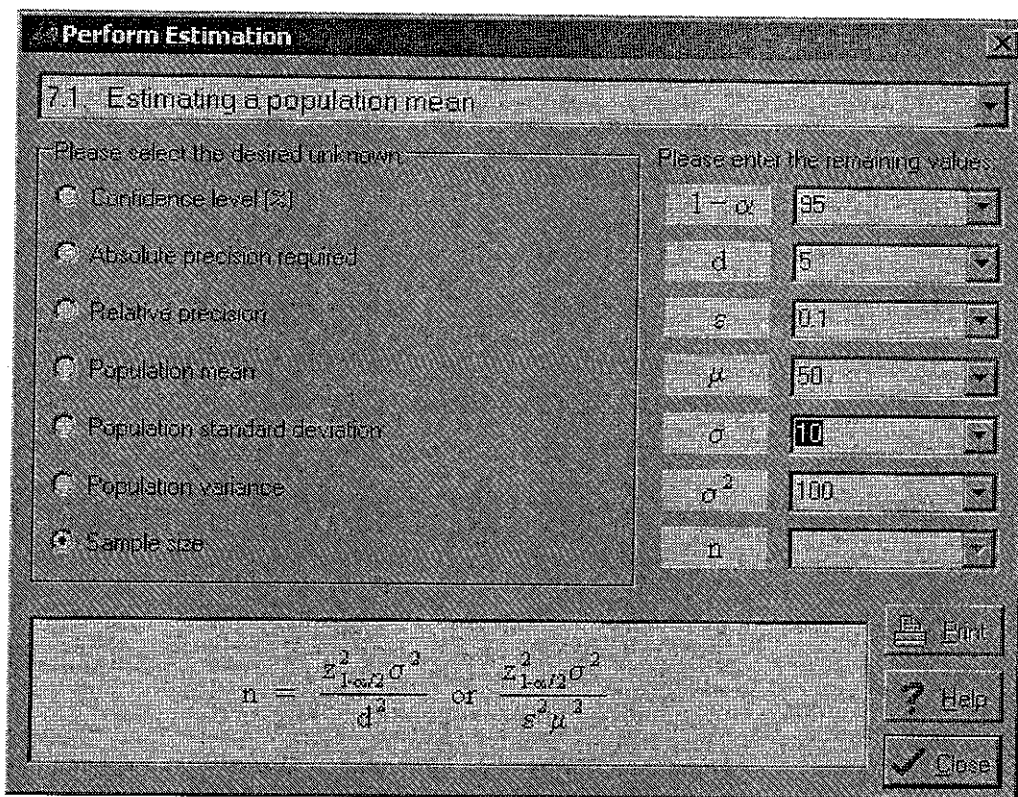
$$n = \frac{z_{1-\alpha/2}^2 \sigma^2}{d^2} \text{ or } \frac{z_{1-\alpha/2}^2 \sigma^2}{\epsilon^2 \mu^2}$$

Print

Help

Close

Đưa 5 vào ô độ tin cậy tuyệt đối (Absolute precision required – d), tương đương với độ tin cậy tương đối là 10%, ($\epsilon = 0,1 = 5/50$). Sau khi các số liệu thích hợp đã được đưa vào hộp cỡ mẫu sẽ được tính một cách tự động.



Phiên bản SSize đưa ra ví dụ đã không tự động tính cỡ mẫu, n. Tuy nhiên nó đưa ra công thức tính và chúng ta có thể tính toán bằng tay, $n = (1,96^2 \times 100)/25 = 16$.

6.3.1.2. Độ tin cậy của ước lượng tỷ lệ

Giả thuyết của câu hỏi số 10 trong bài 3 quan tâm đến ước lượng tỷ lệ toàn bộ quần thể chấn thương giao thông có các chấn thương đầu/chấn thương cột sống, được ước lượng khoảng 37% dựa vào các nghiên cứu trước. Các thống kê mô tả được dự tính dùng để mô tả biến này bao gồm khoảng tin cậy phản ánh độ tin cậy của ước lượng tỷ lệ. Độ tin cậy tuyệt đối yêu cầu là $\pm 10\%$ và tương đương với độ tin cậy tương đối là $10/37 = \pm 27\%$.

H_0 : Tỷ lệ chấn thương ở đầu và cột sống là 37%.

Để xác định cỡ mẫu, dùng công thức độ tin cậy tuyệt đối.

1. Từ màn hình trong phần 6.3, chọn 1.1, nhấp chuột lên nút Estimate. Xuất hiện một màn hình với những hộp còn trống trừ hộp $1-\alpha$ box. Nhập những số phù hợp.

Perform Estimation

1.1. Estimating a population proportion with specified absolute precision

Please select the desired unknown:

- Confidence level (%)
- Anticipated population proportion
- Absolute precision required
- Sample size

Please enter the remaining values:

1 - α 95

P

d

n

$$n = \frac{z_{1-\alpha/2}^2 P(1-P)}{d^2}$$

Print

Help

Close

2. Nhập 0,37 cho tỷ lệ P ước lượng trước của quần thể và 0,10 cho d, độ tin cậy tuyệt đối yêu cầu.

Perform Estimation

1.1. Estimating a population proportion with specified absolute precision

Please select the desired unknown:

- Confidence level (%)
- Anticipated population proportion
- Absolute precision required
- Sample size

Please enter the remaining values:

1 - α 95

P 0,37

d 0,10

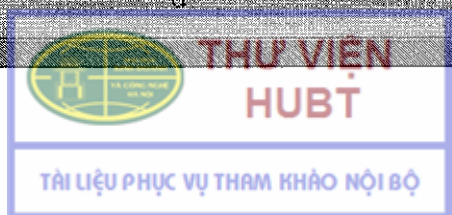
n 90

$$n = \frac{z_{1-\alpha/2}^2 P(1-P)}{d^2}$$

Print

Help

Close



Cỡ mẫu cần thiết cho độ tin cậy này là 90 người.

Để xác định cỡ mẫu cần thiết, sử dụng công thức độ tin cậy tương đối:

1. Từ màn hình trong mục 6.3, chọn 1.2, nhấp chuột lên nút Estimate. Hiện thị một màn hình với những ô còn trống trừ ô $1-\alpha$. Nhập các số liệu như dưới đây.

2. Nhập 0,37 cho tỷ lệ ước lượng P của quần thể, và 0,27 cho ϵ , độ tin cậy tương đối cần thiết. Cỡ mẫu cần thiết để đạt độ tin cậy này là chính xác như đã tính toán ở trên với độ tin cậy tuyệt đối là 10%.

6.3.1.3. So sánh hai trung bình giữa hai nhóm

Giả thuyết của câu hỏi số 5 trong bài 3 có liên quan đến việc so sánh trung bình giữa hai nhóm, kiểm định giả thuyết này là kiểm định rằng không có sự khác biệt về điểm chất lượng cuộc sống trước chấn thương giữa hai nhóm nam và nữ.

H_0 : Điểm trung bình chất lượng cuộc sống của hai nhóm nam và nữ là như nhau.

Trong phạm vi nghiên cứu này, có giả định rằng điểm chất lượng cuộc sống là phân bố chuẩn (dựa trên công cụ là các tài liệu mô tả của một nghiên cứu) và phân bố này trong quần thể chung có trung bình là 50 và độ lệch chuẩn là 10. Trong phần thiết kế nghiên cứu, các nhà nghiên cứu không biết rằng các quần

thể sử dụng loại phương tiện khác nhau khi chẩn thương có phân bố như nhau hay không, nhưng dường như giả định này có vẻ xác thực. Các nhà nghiên cứu muốn đảm bảo rằng họ có đủ nguồn lực tương xứng để tìm ra sự khác nhau có ý nghĩa thống kê về điểm chất lượng cuộc sống là 5 hoặc cao hơn giữa hai giới nam và nữ. Trong kế hoạch nghiên cứu của họ đã quyết định giả thuyết này sẽ được kiểm định bằng kiểm định t không ghép cặp nếu sự khác nhau trung bình giữa nam và nữ bằng 0 (điều này xảy ra khi giả thuyết H_0 là đúng). Vì vậy, dựa trên bảng kiểm 5.4 cho tính cỡ mẫu, họ cần

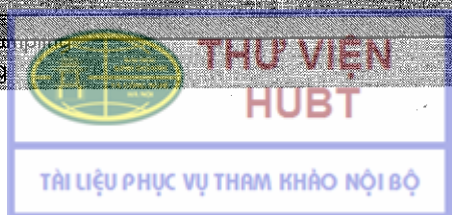
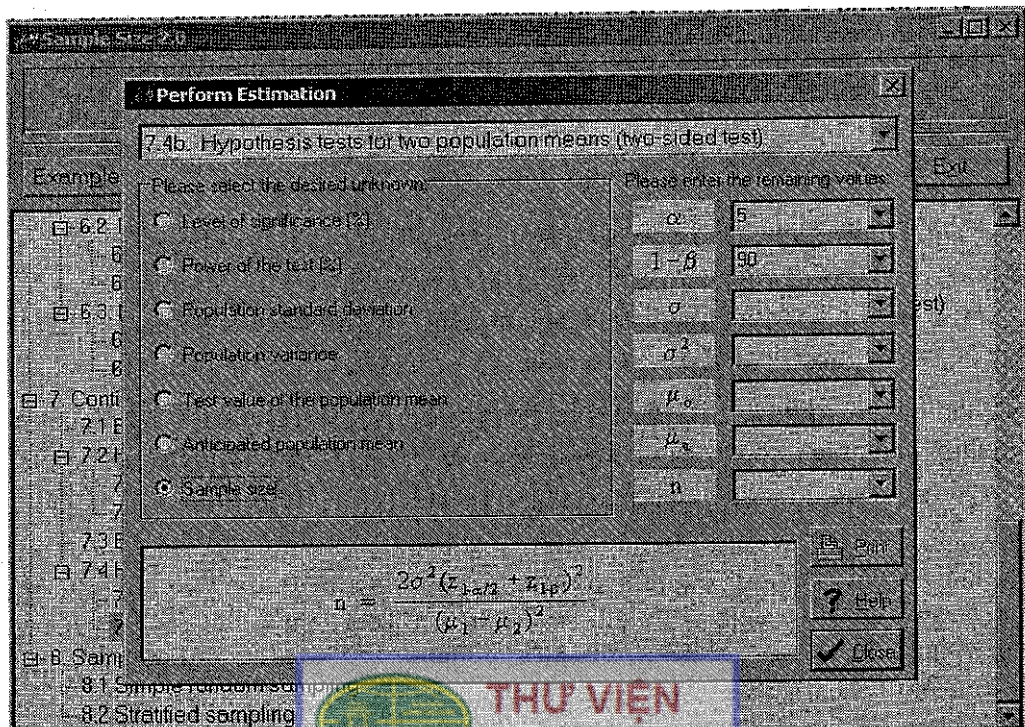
Lượng giá phương sai điểm chất lượng cuộc sống của nam hoặc nữ – điều này đã được giả định là phản ánh độ lệch chuẩn quần thể là 10.

Sự khác nhau nhỏ nhất mà chúng ta quan tâm – sự khác nhau là 5 điểm giữa nam và nữ đã được coi là sự khác nhau tối thiểu có thể tìm ra. Nên nhớ rằng kiểm định so sánh sự khác nhau giữa hai số trung bình sẽ tính toán sự khác biệt giữa 2 trung bình và so sánh nó với giá trị 0.

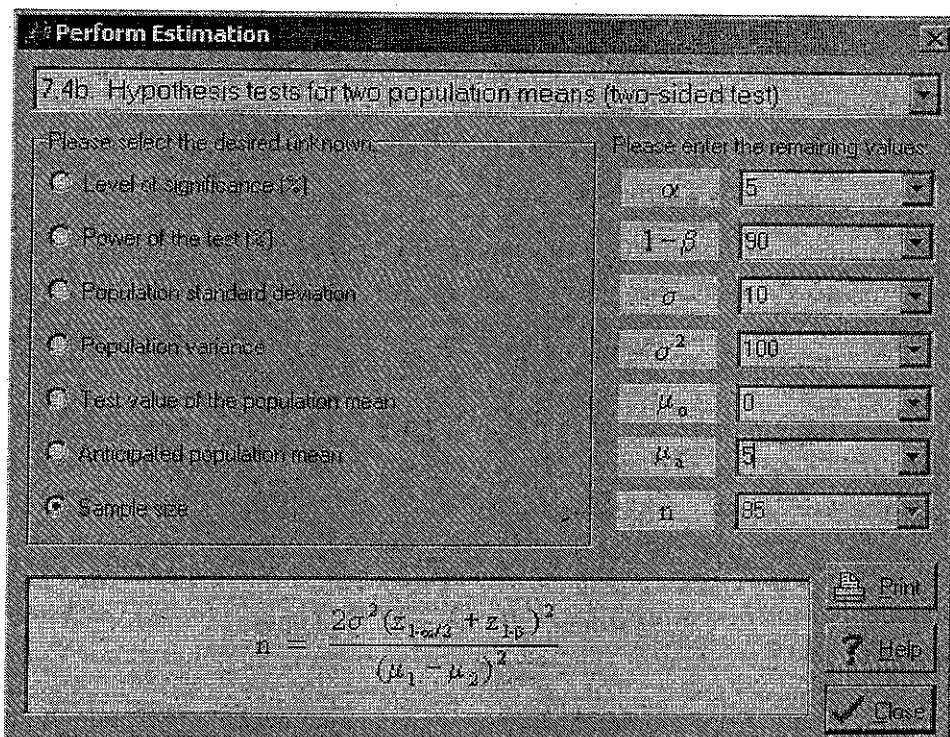
Mức sai lầm loại I có thể mắc phải thường được xác định là 5%. Đó là kiểm định hai phía vì việc giảm giá trị trong bất kỳ nhóm nam và nữ đều được quan tâm. Sai lầm loại II được xác định là 10%, lực kiểm định là 90%.

Để xác định cỡ mẫu cần thiết cho giả thuyết này chúng ta làm như sau:

1. Từ thực đơn như màn hình trong phần 6.3, chọn 7.4b, nhấp chuột lên nút Estimate. Hiện thị màn hình như sau:



2. Đưa giá trị độ lệch chuẩn quần thể là 10 vào, giá trị kiểm định của trung bình quần thể là 0, và trung bình quần thể dự đoán đó chính là sự khác nhau mong đợi giữa trung hai nhóm so sánh (bằng 5). SSize sẽ tự động tính phương sai quần thể và cỡ mẫu.



3. Cỡ mẫu cần thiết cho từng nhóm là 85 người để tìm ra sự khác nhau giữa nam và nữ về điểm chất lượng cuộc sống là 5 hoặc cao hơn và lực kiểm định là 90%.

6.3.1.4. So sánh tỷ lệ giữa hai nhóm

Giả thuyết của câu hỏi 11 trong bài 3 đề cập đến việc so sánh hai tỷ lệ của hai nhóm, kiểm định giả thuyết này là không có sự khác nhau giữa tỷ lệ chấn thương ở đầu/cột sống giữa những người đi bộ và những người sử dụng phương tiện giao thông khác khi bị chấn thương.

H_0 : So với những người sử dụng phương tiện giao thông khi bị chấn thương thì những người đi bộ có tỷ lệ chấn thương ở đầu/cột sống là tương tự hoặc thấp hơn.

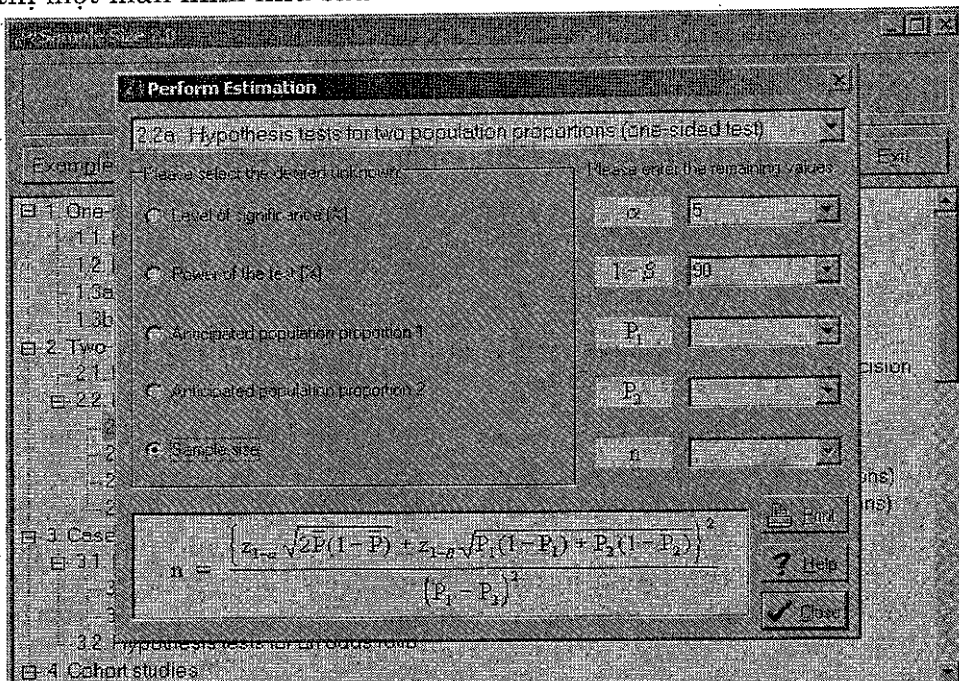
Trong phần thiết kế của nghiên cứu, các nhà nghiên cứu đã biết từ những nghiên cứu chấn thương trước là có khoảng 35% những người bị chấn thương có tổn thương ở đầu/cột sống nếu họ sử dụng phương tiện giao thông. Nhưng họ không biết tỷ lệ này có khác nhau giữa những người đi bộ và những người sử dụng phương tiện hay không. Các nhà nghiên cứu muốn đảm bảo mẫu có đủ lực

để phát hiện ra sự khác nhau giữa các tỷ lệ này là 5% hoặc cao hơn (sự khác nhau tuyệt đối 35% so với 40%). Điều này tương đương với sự khác nhau tương đối là (5/35). Trong kế hoạch phân tích họ đã quyết định kiểm định giả thuyết bằng kiểm định χ^2 mà sẽ kiểm định xem tỷ lệ chung cho tất cả quần thể có sử dụng cho từng nhóm đối tượng chấn thương đi bộ và sử dụng phương tiện không. Vì vậy, để tính cỡ mẫu, họ cần:

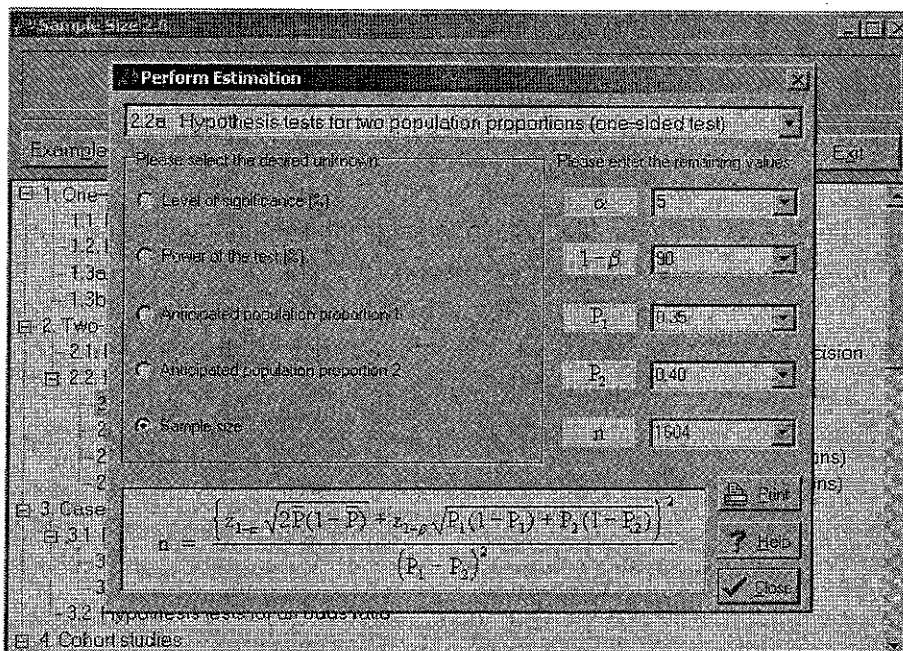
- Lượng giá phương sai là không cần thiết vì đây là so sánh hai tỷ lệ chứ không phải hai số trung bình.
- Sự khác biệt nhỏ nhất – một sự khác biệt 5% (tuyệt đối) giữa chấn thương của người đi bộ và người sử dụng phương tiện được coi là sự khác nhau tối thiểu có thể tìm thấy. Nếu tỷ lệ chấn thương ở đầu/cột sống trong số các đối tượng chấn thương đi xe là 35% thì có nghĩa chúng ta mong đợi tỷ lệ này ở nhóm người đi bộ ít nhất là 40%.
- Mức sai lầm loại I có thể mắc phải thường được xác định là 5%. Đó là kiểm định một phía vì các nhà nghiên cứu chỉ quan tâm đến tỷ lệ chấn thương ở đầu/cột sống ở người đi bộ có cao hơn không. Sai lầm loại II được xác định là 10%, lực kiểm định là 90%.

Xác định cỡ mẫu cần thiết cho giả thuyết này:

1. Từ màn hình trong phần 6.3, chọn 2.2a, nhấp chuột lên nút Estimate. Hiện thị một màn hình như sau



2. Xác định P1 là tỷ lệ chấn thương đầu/cột sống của nhóm sử dụng phương tiện, vì thế đưa 0.35 vào ô P1. P2 là tỷ lệ tỷ lệ chấn thương đầu/cột sống của người đi bộ, đưa 0.4 vào ô P2.



Cỡ mẫu cần thiết cho từng nhóm là 1604 để tìm ra sự khác nhau về tỷ lệ chấn thương ở đầu/cột sống giữa hai nhóm đi bộ và đi xe là 5% hoặc cao hơn với lực kiểm định là 90%. Trên thực tế với trên 1700 người, nghiên cứu NTIS đã không tìm ra sự khác biệt có ý nghĩa thống kê giữa hai nhóm này.

6.3.1.5. Đo lường nhắc lại trong cùng một đơn vị – Giá trị trung bình

Giả thuyết của câu hỏi 6 trong phần 3 đề cập đến kiểm định t ghép cặp, kiểm định giả thuyết mà không có sự thay đổi trong điểm chất lượng cuộc sống sau chấn thương với trước chấn thương.

H₀: Điểm chất lượng cuộc sống sau chấn thương giao thông là cao hơn hoặc không thay đổi so với trước chấn thương.

Các nhà nghiên cứu muốn đảm bảo rằng có đủ lực để tìm ra sự thay đổi có ý nghĩa thống kê về điểm chất lượng cuộc sống sau chấn thương là 5 hoặc cao hơn so với trước chấn thương. Trong kế hoạch phân tích họ đã quyết định giả thuyết này sẽ được kiểm định bằng tính toán một biến mới, sự thay đổi của điểm (trước – sau chấn thương), vì thế họ sử dụng kiểm định t ghép cặp để kiểm tra xem điểm thay đổi trung bình của chất lượng cuộc sống có bằng 0 không (điều này có nghĩa là giả thuyết H₀ là đúng). Vì thế để tính toán cỡ mẫu, họ cần:

Lượng giá phương sai của mức điểm thay đổi – điều này KHÔNG GIỐNG như lượng giá phương sai của điểm chất lượng cuộc sống trước chấn thương hay

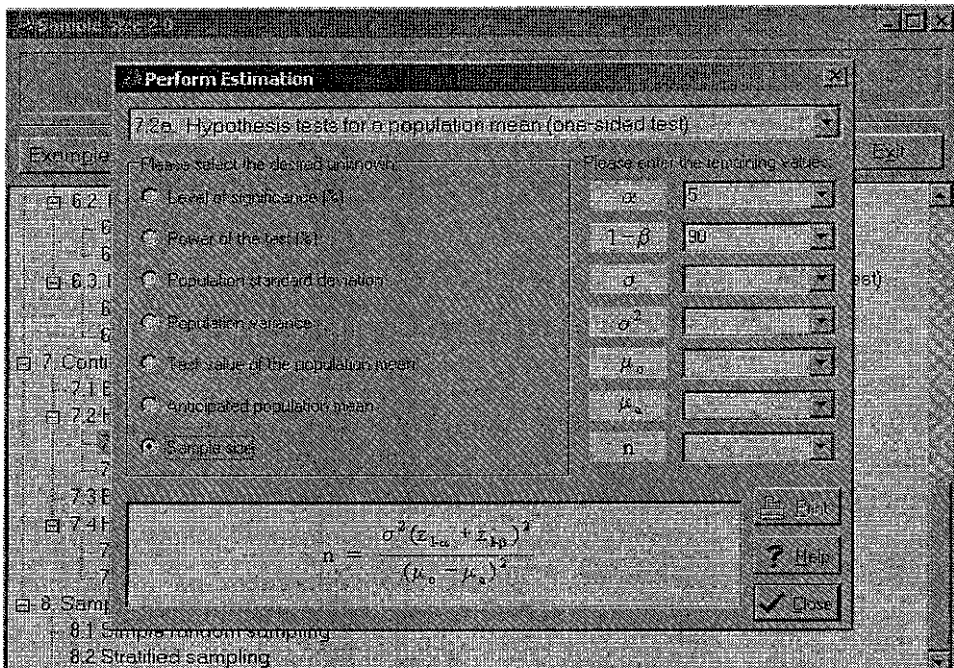
sau chấn thương (điểm thay đổi là một biến khác). Các nhà nghiên cứu cần có vài ý tưởng về độ lệch chuẩn của điểm thay đổi là thế nào. Điều này thường rất khó dự đoán, và thường phải dựa trên các tài liệu nghiên cứu có sẵn. Trong trường hợp nghiên cứu NTIS, do không tìm được các nghiên cứu/tài liệu giúp họ quyết định giá trị này, và vì không có thời gian để thực hiện một nghiên cứu thí điểm nên họ đã đưa ra một giả định là không có một cá nhân nào trong quần thể mà sự thay đổi lại lớn hơn 20 điểm thậm chí ngay cả trong trường hợp chấn thương trầm trọng nhất. Sử dụng giả định này điểm thay đổi sẽ chạy từ -20 đến +20 và có phân bố chuẩn, độ lệch chuẩn của điểm thay đổi được ước lượng chia cho 6 ($40/6 = 6,7^2$).

Sự khác nhau nhỏ nhất được quan tâm – điểm thay đổi có thể tìm ra là 5 điểm.

Mức sai lầm loại I có thể mắc phải thường được xác định là 5%. Đó là kiểm định một phía vì các nhà nghiên cứu chỉ quan tâm đến sự thay đổi là làm giảm chất lượng cuộc sống. Sai lầm loại II được xác định là 10%, lực kiểm định là 90%.

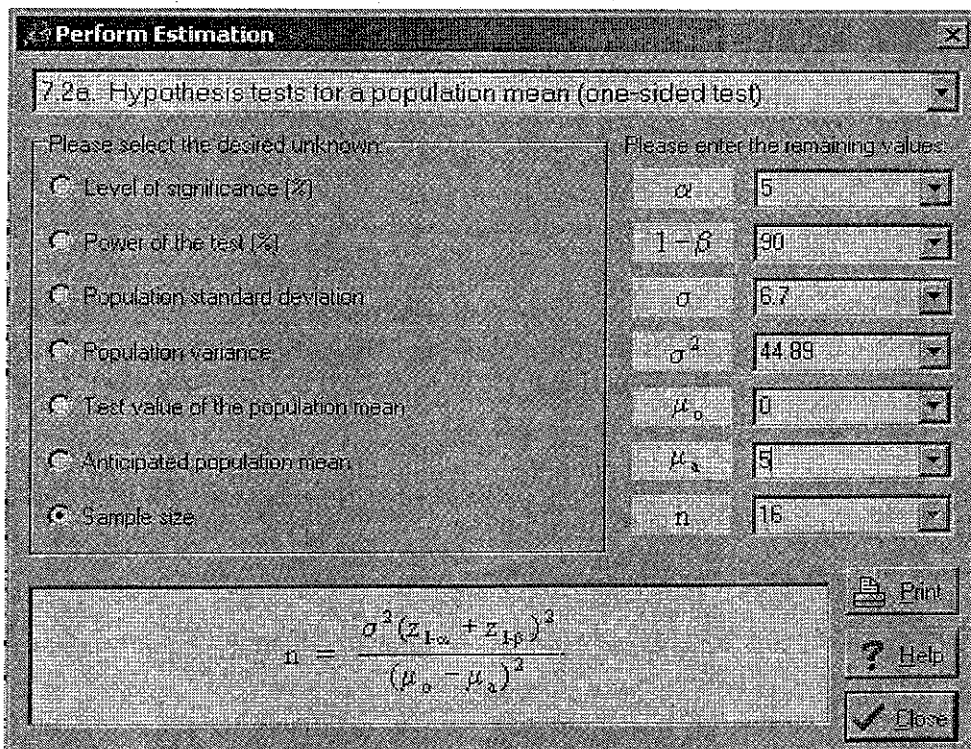
Xác định cỡ mẫu cần thiết cho giả thuyết này như sau:

1. Từ màn hình trong phần 6.3, chọn 7.2a, nhấp chuột vào nút Estimate. Hiện thị một màn hình như sau.



² Với giả định là phân bố chuẩn, do đó sẽ có gần như toàn bộ các giá trị quan sát sẽ nằm trong khoảng $TB+3SD$, vậy khoảng từ giá trị bé nhất tới lớn nhất sẽ là 6 độ lệch chuẩn.

2. Nhập số trung bình quần thể ước lượng trước (sự khác nhau tối thiểu mong đợi có thể tìm ra) là 5, giá trị kiểm định, là 0, và độ lệch chuẩn quần thể là 6,7.



3. Các ô khác sẽ được điền (phương sai và cỡ mẫu). Cỡ mẫu cần thiết là 16 người để tìm ra sự thay đổi điểm chất lượng cuộc sống (trước- sau chấn thương) là 5 điểm.

Hãy nhớ rằng chọn độ lệch chuẩn là 6,7 chỉ là sự suy đoán. Tăng độ lệch chuẩn lên 10.0 cỡ mẫu sẽ tăng lên 35. Vì thế với những giả thuyết đặc biệt như trong nghiên cứu NTIS thì cần một nguồn lực lớn hơn.

6.3.2. Ảnh hưởng của thiết kế nghiên cứu đến cỡ mẫu

Những sự lựa chọn bạn để có thể tìm ra sự khác nhau nhỏ nhất, độ lệch chuẩn, sai lầm loại I và sai lầm loại II sẽ tác động rất lớn đến cỡ mẫu cuối cùng. Tuy nhiên bạn cần nhớ rằng, cỡ mẫu thể hiện số lượng người mà bạn cần thiết phải thu thập số liệu được, trên thực tế không phải lúc nào đối tượng điều tra cũng sẵn sàng trả lời các câu hỏi hoặc bạn sẽ theo dõi được đối tượng trong suốt thời gian nghiên cứu. Vì vậy khi tính cỡ mẫu bạn cũng nên tính đến cả

những trường hợp đối tượng không đáp ứng và tỷ lệ đối tượng bỏ cuộc. Với những cách chọn mẫu phức tạp như mẫu cụm và nhu cầu điều chỉnh các yếu tố nhiễu trong phân tích thống kê mức độ cao hơn cũng sẽ yêu cầu cỡ mẫu lớn hơn những nghiên cứu thực nghiệm hoặc chọn mẫu ngẫu nhiên đơn. Những ảnh hưởng của thiết kế nghiên cứu được trình bày tóm tắt dưới đây.

6.3.2.1. Tác dụng của đường cong lực mẫu

Nếu bạn đang xây dựng kế hoạch thực hiện một nghiên cứu, bạn cần chứng minh cỡ mẫu của bạn là phù hợp. Thường các giá trị bạn chọn cho các tính toán cỡ mẫu chỉ là các ước đoán. Để chắc chắn bạn đang chọn cỡ mẫu phù hợp, bạn nên tính toán cho nhiều trường hợp và sử dụng nhiều giá trị cho sự khác biệt tối thiểu có thể tìm thấy và độ lệch chuẩn. Bạn thể hiện các cỡ mẫu tính được trên một đồ thị, được gọi là đường cong lực mẫu, dựa vào biểu đồ này bạn và đội nghiên cứu sẽ có những quyết định chọn cỡ mẫu nào là có tính khả thi nhất.

Ví dụ, sử dụng những số liệu của nghiên cứu NTIS, ta có điểm trung bình chất lượng cuộc sống là 50 và độ lệch chuẩn là 5, đường cong khả năng cho so sánh giữa hai số trung bình ở trên có dạng:

Đường cong này cho chúng ta thấy các cỡ mẫu tối thiểu cần thiết cho từng nhóm khi giá trị trung bình và độ lệch chuẩn của điểm chất lượng cuộc sống tương ứng là 55 và 10 ở một nhóm. Có thể thấy được trung bình khác nhau tìm được là 10 hoặc hơn (trung bình của nhóm 1 là 55 so với nhóm 2 là 65) thì cần 21 người trong một nhóm để lực kiểm định là 90% hoặc cao hơn nữa. So sánh điều này với cỡ mẫu 84 người trong một nhóm sẽ cho sự khác nhau của điểm QoL nhỏ hơn và bằng 5 (55 đến 60).

Lực kiểm định mạnh hơn thì cỡ mẫu tăng.

Sự khác nhau tối thiểu có thể thấy được nhỏ hơn, cỡ mẫu lớn hơn.

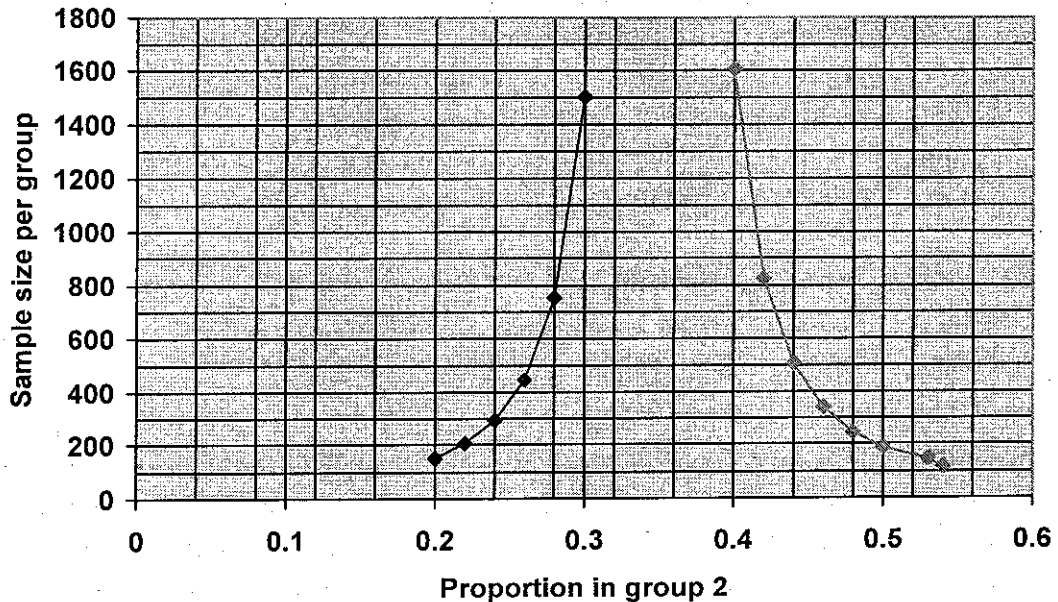
Độ lệch chuẩn lớn hơn, cỡ mẫu lớn hơn. Hãy xem xét đường cong lực mẫu 90% ở trên, độ lệch chuẩn của điểm QoL là 15 hơn là 10.

Cỡ mẫu sẽ lớn hơn khoảng hai lần khi bạn lấy độ lệch chuẩn là 15 so với độ lệch chuẩn là 10. Các tính toán cỡ mẫu là rất nhạy cảm với việc lựa chọn độ lệch chuẩn, vì thế một điều quan trọng là bạn phải có những cơ sở tốt cho sự lựa chọn của mình. Nếu bạn ước lượng độ lệch chuẩn thấp hơn thực tế, bạn có thể làm giảm khả năng kiểm định thống kê của bạn (nghĩa là kết luận so sánh của bạn sẽ có nguy cơ là âm tính giả lớn hơn).

Các đường cong lực mẫu cho các so sánh tỷ lệ rất khác với những đường cong so sánh các số trung bình trên khi tỷ lệ kiểm định rất lớn hoặc rất nhỏ. Khi các tỷ lệ này gần 50%, đường cong khả năng có hình dạng tương tự như so

sánh các số trung bình. Ví dụ, hãy xem xét đường cong khả năng cỡ mẫu cho so sánh của những người đi bộ bị chấn thương ở đầu/cột sống với những người đi xe.

Power curves for prevalence of head & spinal injury in pedestrian versus vehicle accidents
(proportion group 1 = 35%, power=90%, one-tailed)



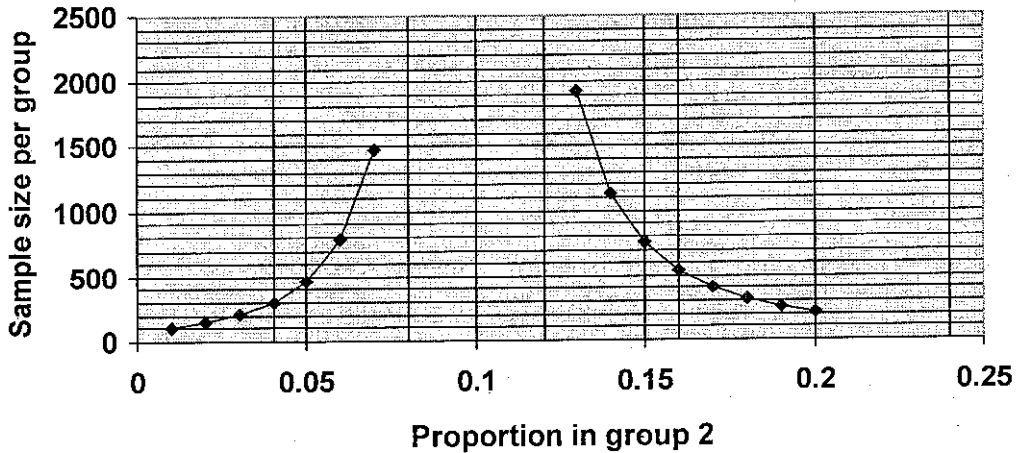
Những điểm cần lưu ý:

- Cỡ mẫu cần thiết để đưa ra sự khác biệt về tỷ lệ lớn hơn phân tích sự khác biệt về trung bình.
- Cỡ mẫu càng tăng khi tìm kiếm sự khác biệt nhỏ.

Không như đường cong lực mẫu khi so sánh các giá trị trung bình, các đường cong này không đối xứng – Cỡ mẫu cần thiết để phát hiện tỷ lệ tăng (35% lên 40% = +5%, màu hồng) không giống với cỡ mẫu phát hiện tỷ lệ giảm với một mức tương tự (35% xuống 30% = -5%, màu xanh).

Biểu đồ sau trình bày các đường cong lực mẫu khi các hiện tượng quan tâm ít thông dụng hơn, khoảng 10%. Tính không đối xứng càng được thể hiện rõ trong trường hợp này.

Power curves for prevalence of head & spinal injury in pedestrian versus vehicle accidents (proportion group 1 = 10%, power=90%, one-tailed)

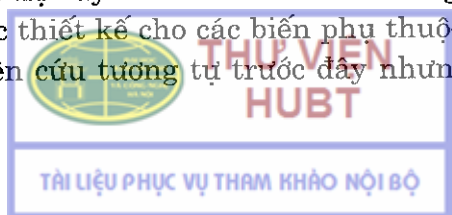


6.3.2.2. Các yếu tố điều chỉnh sự ảnh hưởng của thiết kế

Với việc sử dụng các thiết kế phức tạp hơn, cỡ mẫu sẽ cần phải lớn hơn để đạt được cùng một lực kiểm định tìm ra sự khác nhau cần thiết. Bất kỳ thiết kế nào phức tạp hơn thử nghiệm ngẫu nhiên sử dụng phương pháp chọn mẫu ngẫu nhiên đơn cũng sẽ cần đến một vài giá trị để điều chỉnh cỡ mẫu khi sử dụng phần mềm SSize.

Giá trị mẫu tối thiểu cần phải được cộng thêm một giá trị nào đó để điều chỉnh các vấn đề:

Nếu cách lấy mẫu cụm hoặc mẫu nhiều giai đoạn được sử dụng, tăng cỡ mẫu lên từ 1,6 – 3,0 lần phụ thuộc vào tính tương đồng mà bạn mong đợi (cho biến phụ thuộc) trong cụm. Ví dụ, sức khỏe tâm thần của hai đứa trẻ trong cùng một trường học dường như sẽ không có mối tương quan nhiều như là hai đứa trẻ từ hai trường khác nhau. Trường học là một biến cụm, nhưng sức khỏe tâm thần của trẻ chịu ảnh hưởng của rất nhiều yếu tố bên ngoài hơn là những yếu tố trong trường học. Ngược lại, nếu biến phụ thuộc là kết quả học tập môn toán thì điểm toán của hai đứa trẻ trong cùng một trường lại có tương quan lớn hơn điểm của hai đứa trẻ ở hai trường khác nhau (nếu việc giảng dạy là yếu tố chính quyết định việc học tập). Yếu tố hiệu chỉnh sự co cụm được biết đến là giá trị hiệu lực thiết kế, và giá trị này sẽ khác nhau khi chúng ta xem xét các biến độc lập khác nhau. Hiệu lực thiết kế cho các biến phụ thuộc của bạn có thể đã được công bố trong các nghiên cứu tương tự trước đây nhưng nhìn chung là rất khó



tìm được và hầu hết mọi trường hợp bạn sẽ phải tự ước đoán giá trị này. Bạn sẽ không biết chắc chắn là hiệu lực thiết kế mạnh thế nào cho đến khi bạn hoàn thành nghiên cứu, trên thực tế khi thiếu thông tin về hiệu quả thiết kế các nhà nghiên cứu vẫn dùng một giá trị ngầm định bằng 2.0.

Hãy cân nhắc tỷ lệ đối tượng trong mẫu của bạn sẽ bỏ cuộc (nếu bạn có một thiết kế nghiên cứu dọc). Ví dụ, hãy tăng cỡ mẫu của bạn lên 1,2 nếu bạn nghĩ có thể có 20% người bỏ cuộc trong nghiên cứu của bạn.

Nếu thiết kế nghiên cứu của bạn không phải là nghiên cứu thực nghiệm, bạn sẽ phải điều chỉnh nhiều bằng cách dùng phương pháp thống kê phức tạp hơn. Tăng mẫu lên 1,2 để đáp ứng yêu cầu này.

Cuối cùng, cân nhắc đến tỷ lệ người tham gia/đáp ứng bạn sẽ đạt được. Ví dụ, nếu bạn mong muốn đạt được 70% tham gia, bạn sẽ cần tăng mẫu lên 1,4 (= 100%/70%).

Bạn có thể thấy cỡ mẫu tối thiểu là 50 cho từng nhóm có thể nhanh chóng trở thành 202 cho mỗi nhóm nếu tất cả các yếu tố trên được phù hợp!!!

Cân nhắc cẩn thận những thiết kế có thể có tác động lên cỡ mẫu, từ các giả định bạn sử dụng với mỗi công thức tới những vấn đề phát sinh mà cần phải quan tâm đến. Những gợi ý ở đây sẽ cho bạn một số ý tưởng về cỡ mẫu cần cho nghiên cứu của bạn, tuy nhiên nên tham khảo ý kiến chuyên gia thống kê khi bạn tính cỡ mẫu trong trường hợp các thiết kế nghiên cứu phức tạp.

CHỌN LOẠI KIỂM ĐỊNH THỐNG KÊ THÍCH HỢP – KHI BIẾN ĐẦU RA LÀ LIÊN TỤC HOẶC KHOẢNG

		Phân bố của biến đầu ra có tuân theo phân phối chuẩn hay không ?					
		Có			Không		
		Bao nhiêu nhóm sẽ được so sánh?					
		1 ^a	2	3+	Bao nhiêu nhóm sẽ được so sánh?		
		1 ^a	1 ^b	1 ^c	1 ^d	2	
		Đầu ra là liên tục	So sánh với giá trị có sẵn	Đo lường lặp lại trên cùng một đơn vị	So sánh với giá trị có sẵn	Đo lường lặp lại trên cùng một đơn vị	
Các thống kê mô tả		Hệ số tương quan Pearson	TB & ĐLC	TB sự khác biệt & ĐLC ^a của sự khác biệt	TB & ĐLC	TB & ĐLC	Các TV & Khoảng
Các kiểm định thống kê		Kiểm định tương quan	KD t một mẫu	KD t-ghép cặp	KD t-không ghép cặp	KD phân tích phương sai (ANOVA)	Kiểm định Mann-Whitney
Giá thuyết (H ₀)		Tương quan = 0	TB = giá trị chuẩn	Sự khác biệt giữa hai TB = 0	TB 1 = TB 2	TB1 = TB2 = TB3= .v.v.	Kiểm định Kruskal-Wallis ANOVA
Các giá định		Hai biến có quan hệ tuyến tính	Biến đầu ra tuân theo phân phối chuẩn	Sự khác biệt tuân theo phân phối chuẩn	Biến đầu ra tuân theo phân phối chuẩn và phương sai đồng nhất	Biến đầu ra tuân theo phân phối chuẩn và phương sai đồng nhất	Kiểm định TV1 = TV2 = TV3 = .v.v.

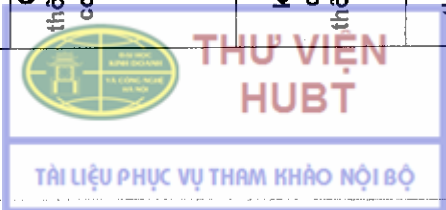
- a) Đây không phải là nhóm mà chỉ là một biến độc lập liên tục
- b) Đây không phải là biến độc lập mà chỉ là một thống kê mô tả mà được so sánh với một quần thể chuẩn hay giá trị ước tính
- c) Là thiết kế nghiên cứu trong đó nhiều đo lường được tiến hành lặp lại trên cùng một đơn vị quan sát
- d) Không lẫn với kiểm định tổng hạng Wilcoxon (tương tự như kiểm định Mann-Whitney)
- e) Viết tắt: ĐLC - Độ lệch chuẩn; TB - Trung bình; TV - Trung vị; KD - Kiểm định

CHỌN LOẠI KIỂM ĐỊNH THÍCH HỢP – KHI BIẾN ĐẦU RA LÀ BIẾN PHÂN LOẠI

Có bao nhiêu phân nhóm trong biến đầu ra

		2				3+				
		Độ đo các biến được so sánh như thế nào				Độ đo của các biến so sánh như thế nào?				
	Phân loại - 2 phân nhóm	Phân loại - từ 3 phân nhóm trở lên	Liên tục hoặc khoảng - phân bố chuẩn	Liên tục hoặc khoảng - phân bố chuẩn	Phân loại lập lại trong một đơn vị quan sát	phân loại - 2 phân nhóm	Phân loại - từ 3 phân nhóm trở lên	Liên tục hoặc khoảng - phân phối chuẩn	Liên tục hoặc khoảng - phân phối chuẩn	Các loại được lặp lại (trong một đơn vị quan sát)
Các thống kê cơ bản	% hoặc OR^a	% hoặc OR	TB & ĐLC của các biến liên tục trong các phân nhóm của biến đầu ra	TV và khoảng của các biến liên tục trong các phân nhóm của biến đầu ra	% phù hợp, Hệ số kappa về tính phù hợp và sai số chuẩn	%	%	TB & ĐLC của biến liên tục trong các phân nhóm của biến đầu ra	TV & khoảng của biến liên tục trong các phân nhóm của biến đầu ra	%
Các kiểm định thống kê	Kiểm định χ^2	Kiểm định χ^2	KĐ t không ghép cặp	Kiểm định Mann-Whitney	Kiểm định McNemar hoặc kiểm định kappa cho mức ý nghĩa	Kiểm định χ^2	Kiểm định χ^2	Phân tích phương sai (ANOVA)	Kruskal-Wallis ANOVA	Kiểm định kappa cho mức ý nghĩa
Giá thuyết (H_0)	%1 = %2 hoặc OR = 1	tất cả % bằng nhau hoặc tất cả $OR^a = 1$	TB1 = TB2	TV1 = TV 2	Sự không phù hợp của các phân nhóm cùng hướng	Tần số các ô quan sát gọi ý rằng hai biến là độc lập	Tần số các ô quan sát là độc lập	tất cả các TB là tương tự nhau	tất cả TV tương tự nhau	Sự không phù hợp của các phân nhóm cùng hướng
Các giá định	Các quan sát là độc lập	Các quan sát là độc lập	Biến đầu ra tuân theo phân phối chuẩn	Biến đầu ra tuân theo phân phối chuẩn; các phương sai đồng nhất	Các đơn vị quan sát là độc lập	Các quan sát là độc lập	Các quan sát là độc lập	Biến đầu ra tuân theo phân phối chuẩn	Biến đầu ra tuân theo phân phối chuẩn; các phương sai đồng nhất	Các đơn vị quan sát là độc lập

Viết tắt: OR - Tỷ số chênh; ĐLC - Độ lệch chuẩn; TB - Trung bình; TV - Trung vị; KĐ - Kiểm định



NHÀ XUẤT BẢN Y HỌC

THỐNG KÊ Y TẾ CÔNG CỘNG

Phân Phân Tích Số Liệu

Chịu trách nhiệm xuất bản

HOÀNG TRỌNG QUANG

<i>Biên tập:</i>	BÙI THỊ TRÀ GIANG
<i>Sửa bản in:</i>	BÙI THỊ TRÀ GIANG
<i>Trình bày bìa:</i>	CHU HÙNG
<i>Kt vi tính:</i>	BÙI THỊ TRÀ GIANG



In 1000 cuốn, khổ 19x27 tại Xưởng in Nhà xuất bản Y học.

Giấy phép xuất bản số: 603 - 2009/CXB/2 - 62/YH

In xong và nộp lưu chiểu quý III năm 2009.

TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ