Bài giảng 4:
Quản lí dữ liệu

Nguyễn Văn Tuấn

Viện nghiên cứu y khoa Garvan, Australia

Giáo sư thỉnh giảng, Đại học Tôn Đức Thắng, Việt Nam

OMB # 0925-0334 Exp. 6/97

Cardiovascular Health Study

SEVENTH FOLLOW-UP (YEAR 9)
INFORMANT QUESTIONNAIRE ON COGNITIVE DECLINE

Name
ID#:

INSTRUCTIONS: We want you to remember what your friend or relative was like 10 years ago and to compare it with what he/she is like now. Ten years ago was in 1985. Below are situations in which this person has to use his/her memory or intelligence. We want you to tell us whether this has improved, stayed the same, or become worse in the following situations during the past 10 years. It is importance to compare his/her present performance with ten years ago. So, if ten years ago this person always forgot where he/she left things and he/she still does, you would mark, "not much change." Please check the appropriate answer for each item to the best of your knowledge.

Compared with ten years ago how is this person at:

	Much improved	A bit improved	Not much change	A bit worse	Much worse	Don't Know
Recognizing the faces of friends.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 9
Remembering the names of family and friends.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 9
Remember things about family and friends, such as their occupations, birthdays, and addresses.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 9
Remembering things that happened recently	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 9
Recalling conversations a few days later.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 9
Forgetting what he/she wanted to say in the middle of a conversation.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 9
Remembering her/his address and telephone number.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 9
Remembering what is the day and month.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 9
Remembering where things are usually kept.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 9
Remembering where to find things that have been put in a different place than usual.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 9
Adjusting to any change in his/her daily routine.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 9
Knowing how to work familiar machines around the house.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 9
Learning to use a new gadget or machine around the house.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 9
Learning new things in general.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 9

Dữ liệu gốc (thô)

Data: Likert scale

- Likert scale: dùng để đánh giá mức độ đồng thuận của một phát biểu / sự kiện

Mức độ mà bạn đồng ý hay không đồng như với phát biểu sau đây: ....

- ☐ Rất đồng ý
- ☐ Đồng ý
- ☐ Trung dung
- ☐ Không đồng ý
- ☐ Rất không đồng ý

Dữ liệu (data)

- Dữ liệu khoa học là **vàng**, là **kim cương**
- Một phần rất quan trọng của nghiên cứu khoa học
- Là chứng từ của nghiên cứu
- Có thể sử dụng *nhiều lần* sau này
- Có thể phải chia sẻ với đồng nghiệp quốc tế (data sharing)
- Ý nghĩa đạo đức khoa học

Hai loại dữ liệu

- Bản gốc (giấy)
- Bản điện tử
- (Có thể kể đến một số output)

Phần mềm để lưu trữ dữ liệu điện tử

- Microsoft Access
- Epi Info
- Excel
- Oracle

Có khi nhập dữ liệu 2 lần

Nguyên tắc chuẩn bị dữ liệu cho phân tích

- **Nguyên tắc 1: Dòng và cột** (dòng là *quan sát,* cột là *biến số*)
- **Nguyên tắc 2: Mỗi biến là mỗi cột**
- **Nguyên tắc 3: Tất cả cột phải có số liệu**, kể cả missing data
- **Nguyên tắc 4: Nhập dữ liệu gốc**, không phải tính toán
- **Nguyên tắc 5: Dữ liệu trống (missing data)** phải được mã hoá thích hợp

Nguyên tắc 1: dòng = observation, cột=variable

- Tất cả các chương trình máy tính dùng cho phân tích đều dùng dữ liệu theo dạng tabular hay ma trận (dòng và cột)
- Với các dữ liệu điều tra, cột thường thể hiện một biến đơn hay một câu hỏi, dòng thể hiện id của đối tượng

	Id	Age	Gender	Service	employed
Case 1	1	27	1	2	1
Case 2	2	19	2	1	2
Case 3	3	24	2	3	1

Nguyên tắc 2: mỗi biến là một cột

SubID	Time	Response
ADJ	1	183
ADJ	2	177
ADJ	3	192
BDR	1	186
BDR	2	183
BDR	3	169

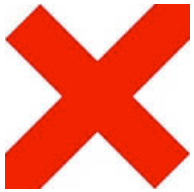
SubID	Time 1 Response	Time 2 Response	Time 3 Response
ADJ	183	177	192
BDR	186	183	169

SubID	Time	Response	SubID	Time	Response
ADJ	1	183	BDR	1	186
ADJ	2	177	BDR	2	183
ADJ	3	192	BDR	3	169

Nguyên tắc 3: Cột phải có số liệu!

SubID	Time	Response
ADJ	1	183
ADJ	2	177
ADJ	3	192
BDR	1	186
BDR	2	183
BDR	3	169

SubID	Time	Response
ADJ	1	183
	2	177
	3	192
BDR	1	186
	2	183
	3	169



Khi có nhiều files ...

- Nếu có nhiều file, không nên dựa vào tên của file để lưu trữ thông tin.
- Thay vì dùng nhiều files, có thể chỉ cần thêm cột để chỉ thông tin mới.

Filename: Anjou

SubID	Time	Response	HeartRate
ADJ	1	183	120
ADJ	2	177	115
ADJ	3	192	101

Filename: Bartlett

SubID	Time	Response	HeartRate
BDR	1	186	112
BDR	2	183	115
BDR	3	169	135

Filename: Anjou

SubID	Time	Response	HeartRate
ADJ	1	183	120
ADJ	2	177	115
ADJ	3	192	101

Filename: Bartlett

SubID	Time	Response	HeartRate
BDR	1	186	112
BDR	2	183	115
BDR	3	169	135

SubjID	Time	Response	HeartRate
ADJ	1	183	120
ADJ	2	177	115
ADJ	3	192	101
BDR	1	186	112
BDR	2	183	115
BDR	3	169	135

Dữ liệu polytomous

Nếu dữ liệu có nhiều nhóm hay classes, không bao giờ giảm xuống phần trăm hay tỉ lệ, mà phải nhập dữ liệu gốc:

Red	White	Blue
10	25	2
5	50	1

Red	White	Blue
27.0	67.5	5.4
8.9	89.3	1.8

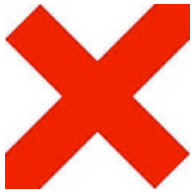
61- Ấn tượng -08	62- Ấn tượng -09	63- Ấn tượng -10	64- Ngành học	65- Lý do chọn ngành-Dư luận-01	66- Lý do chọn ngành-Dư luận-02	67- Lý do chọn ngành:triển vọng thu nhập-03	68- Lý do chọn ngành:triển vọng thu nhập-04	69- Lý do chọn ngành-05	70- Lý do chọn ngành-06	71- Lý do chọn ngành-07	72- Lý do chọn ngành-08	73- Lý do chọn ngành-09	74- Lý do chọn ngành-10
1	0	0	Điện - Điện tử	1	0	1	0	0	0	1	0	0	0
1	0	1	Điện - Điện tử	1	0	1	1	1	0	1	1	1	1
0	0	0	Điện - Điện tử	1	0	0	1	0	0	1	0	0	0
0	1	0		1	0	0	1	0	0	1	0	0	0
1	1	0	Kỹ thuật Điện - Điện tử	0	1		1	0	0	1	1	0	0
1	0	0	Điện - Điện tử	1	0		1	1	0	1	0	0	0
0	1	0	Điện - Điện tử	0	1	1	0	0	0	1	1	0	0
0	0	0		1	0	1	0	0	0	1	0	1	0
0	1	0		1	1	1	0	1	0	0	0	1	0
0	0	0		0	0	1	0	0	0	1	0	0	0
1	0	0		1	0	1	1	0	0	0	1	0	0
1	0	0		1	0	1	0	0	0	1	0	0	0
1	0	0	Điện tử	1	0	0	1	1	0	1	0	0	0
1	1	0		0	0	0	0	0	0	0	0	0	0
0	0	0	điện	1	0	1	0	0	0	0	1	0	1

Nguyên tắc 3: Nhập dữ liệu gốc

- Không nhập tỉ lệ, mà chỉ nhập tử số và mẫu số
- Tử số và mẫu số cần phải có cột riêng để dễ tính toán

Dose	Total	Dead
1	90	5
2	85	30
3	93	60

Dose	% Dead
1	5.5
2	35.3
3	64.5



Nguyên tắc 5: Missing data phải mã hoá

- Tất cả các cột phải có cùng số dòng (kể cả missing data (số khống).
- **Dùng "blank space" hoặc "." hoặc "NA" để chỉ missing data.**

  - Tuyệt đối không dùng 0 hay 999 cho missing data!

Tạo dữ liệu: những điều quan trọng cần biết

- Không dùng header, trailer, subtotals, hay những thông tin "ngoại biên"
- Tên biến số có ý nghĩa và dễ đọc

			XET NGHIEM																
Nguồn dữ liệu						phiếu thu							phiếu thu						
Năm	STT	Số BA	**XN sinh hóa phiếu thu**	XN sinh hóa bệnh án	XN sinh hóa	**glucoza máu**	**HbA1c**	**Máu lắng**	**XN huyết học phiếu thu**	XN huyết học bệnh án	XN huyết học	**Khí máu**	**Định nhóm máu**	**đường giấy**	**anti HCV**	**HBs Ag**	**XN Cross - Match**	**test coombs**	**procalcitonin**
2015	1	38	**1**		1			**1**	**1**	1	1		**1**		**1**	**1**			
2014	166	202	1	2	2	1	0	0	2	3	3	0	2	0	0	0	1	1	
2015	2	289	4	5	5	4	1	1	2	3	3	1	2	0	1	1	1	2	
2015	3	382	1	-99	1	2	0	1	0	1	1	0	0	0	1	1	0	0	
2015	4	409	3	4	4	3	0	0	0	5	5	0	3	0	1	1	2	2	
2015	5	460	0	1	1	1	1	1	1	2	2	1	0	0	1	1	0	0	
2014	167	478	-99	3	3	-99	-99	1	1	2	2		4		1	1	4	2	
2015	6	568	0	1	1	0	0	1	0	1	1	0	0	0	1	1	0	0	
2015	7	613	0	2	2	1	1	1	1	5	5	2	0	0	1	1	0	0	
2015	8	629	2	1	2	1	1	0	0	1	1	1	5	0	1	1	4	4	
2014	168	691	0	2	2	1	0	1	0	4	4	0	0	0	1	1	0	0	
2015	9	759	1	1	1	2	0	1	3	1	3	1	0	0	1	1	2	2	
2014	169	871	0	1	1	1	0	1	0	3	3	0	0	0	1	1	0	0	
2014	170	949	0	1	1	0	0	1	2	1	2	0	0	0	1	1	2	2	

Chuẩn bị dữ liệu tốn thời gian

Qui luật chung:

- 90% chuẩn bị dữ liệu
- 10% phân tích

Phải hết sức cẩn thận với dữ liệu!

Áp dụng qui tắc phòng lab

- Qui tắc lab: Có sổ ghi dữ liệu (red book)
- Mỗi khi thay đổi số liệu, phải có ghi chú và giải thích
- Trong phân tích dữ liệu cũng có qui tắc
  - Tất cả sửa đổi phải có chú thích + giải thích + kí tên
  - Mã hoá (coding) phải có chú tích
  - Dùng LabArchive (nếu có)

Tóm lại: 5 nguyên tắc

- **Nguyên tắc 1: Dòng và cột** (dòng là *quan sát,* cột là *biến số*)
- **Nguyên tắc 2: Mỗi biến là mỗi cột**
- **Nguyên tắc 3: Tất cả cột phải có số liệu**, kể cả missing data
- **Nguyên tắc 4: Nhập dữ liệu gốc**, không phải tính toán
- **Nguyên tắc 5: Dữ liệu trống (missing data)** phải được mã hoá thích hợp