INSTITUTE OF SOCIAL AND MEDICAL STUDIES

CHUẨN BỊ SỐ LIỆU CHO PHÂN TÍCH

Nguyễn Trương Nam
Nguyễn Thị Trang

Nội dung

- Làm sạch số liệu
- Tạo biến mới (sử dụng SPSS)
- Kiểm tra phân bố chuẩn (sử dụng SPSS)

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Làm sạch số liệu (1)

- **Lỗi khi thu thập số liệu:**
  - Lỗi chuyển câu,
  - ghi lại đáp án rất khó nhìn hay được phiên dịch sang nghĩa khác…
  - Lỗi ngoài khoảng cho phép
  - Missing
- **Lỗi Nhập liệu:**
  - Lỗi nhập đáp án,
  - Nhập số liệu ngoài khoảng cho phép,
  - Bỏ sót đáp án, ..

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Làm sạch số liệu (2)

- Hạn chế lỗi Khi thu thập số liệu:
  - Huấn luyện DTV
  - Giám sát chất lượng
  - Kiểm tra phiếu, sửa lỗi tại thực địa
- Khi nhập liệu
  - Viết các lệnh logic, lệnh đặt khoảng cho phép (range), bắt buộc nhập…
  - Huấn luyện nhập liệu viên,
  - Nhập liệu 2 lần và so sánh (validation)

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Làm sạch số liệu (3)

- **Kiểm tra số liệu**
  - **Lệnh Descriptives**: giúp xác định được giá trị lớn nhất, giá trị nhỏ nhất cho các biến, giúp dễ dàng nhận thấy những giá trị được nhập ngoài khoảng cho phép.
  - Lệnh Descriptives cũng giúp xác định được giá trị trung bình cho các biến liên tục, từ đó có thể xác định được những bất thường xảy ra.
  - **Lệnh Frequency**: để kiểm tra value labels và các giá trị bất thường, kiểm tra số lượng trường hợp mất thông tin cho từng biến.
  - **Lệnh Sort case** để xem các giá trị bất thường.
  - **Viết lệnh logic check:**
  - Kiểm tra lỗi do chuyển câu: Nên viết lệnh bằng syntax để kiểm tra số liệu với những bộ câu hỏi có những bước nhảy phức tạp (có thể đã được kiểm tra qua logic checking ở Epidata).
  - Kiểm tra giá trị logic giữa các biến.

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Ví dụ kiểm tra các biến bằng lệnh mô tả và lệnh tần suất

- Tuổi (q2).
- Dân tộc (Q4).
- Trình độ học vấn (Q5).
- Tình trạng hôn nhân(Q7).
- Bạo lực tinh thần (Q113)
- Số lần nạo phá thai (Q40).

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Ví dụ kiểm tra các biến bằng lệnh mô tả và lệnh tần suất

- DESCRIPTIVES VARIABLES=q2

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Q2- Number of years	1277	32	17	49	35.93	8.151
Valid N (listwise)	1277					

Ví dụ kiểm tra các biến bằng lệnh mô tả và lệnh tần suất

- FREQUENCIES VARIABLES=q4

Statistics

Q4-Religion

N	Valid	1281
	Missing	0

Q4-Religion

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	None	1216	94.9	94.9	94.9
	Buddhism	16	1.2	1.2	96.2
	Christian Catholic	48	3.7	3.7	99.9
	Other	1	.1	.1	100.0
	Total	1281	100.0	100.0	

ISMS INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Làm sạch số liệu (4)

- Ví dụ: Kiểm tra lỗi bước nhảy giữa câu Q26 (đã bao giờ quan hệ tình dục chưa? Nếu 0-Sẽ không hỏi câu Chị đã từng mang thai chưa? (Q31).

  compute clean1=0.

  If (q26=0 and ((q31=1) or (q31=0))) clean1=1.

  fre var clean1.

Nếu những phiếu có bước nhảy sai thì clean1=1.

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Làm sạch số liệu (5)

- Ví dụ: Kiểm tra logic giữa câu: Q39: Trong vòng 5 năm qua chị xảy thai bao nhiêu lần? và Q38-Từ trước đến nay chị xảy thai bao nhiêu lần? Q39<=Q38.

compute clean2=0.

if (q39 gt q38) clean2=1.

fre var clean2.

Những phiếu sai logic thì clean2=1. (Phiếu:42416-Q38=0, Q39=1).

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

MỘT SỐ LỆNH TẠO BIẾN TRONG SPSS

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Một số lệnh cơ bản của SPSS

- **Select** case (Lựa chọn các trường hợp)
- **Recode** into the same variable (Mã hóa lại biến mới thay thế biến cũ)
- **Recode** into different variable (Mã hóa lại biến mới giữ nguyên biến cũ)
- **Compute** variables (Tạo các biến mới)

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Ví dụ.

- Sử dụng một số biến có trong bộ số liệu nghiên cứu tại Thái Nguyên (bộ câu hỏi được upload theo bài giảng).
- Tuổi (q2).
- Dân tộc (Q4).
- Trình độ học vấn (Q5).
- Tình trạng hôn nhân(Q7).
- Bạo lực tinh thần (Q113)
- Số lần nạo phá thai (Q40).

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Lệnh ví dụ

- Chọn những người dân tộc kinh (Select cases)

  Data/select cases/if/Q3=1.

- Recode lại biến số lần nạo phá thai thành biến nhị phân 1 “Nạo phá thai lặp lại>=2 lần” 0 “Không nạo phá thai lặp lại<=1 lần” (Recode)

  Recode q40 (2 thr HIGHEST=1) (0=0) (1=0) (missing=sysmis) into Q40_re_abor.

  VARIABLE LABELS Q40_re_abor "Q40-repeated abortion in the life time".

  value labels q40_re_abor 1 "repeated_abortion_life“ 0 “No repeated abortion in lifetime”.

  execute.

- Tạo một biến có bị bạo lực về emotional từ câu hỏi 113. (Lệnh compute).

- *****QUESTION 113 - emotional violence.*
- *********Ever had emotional violence*

- *compute GBV_emo_e=9.*
- *if (b113a1=1) or (b113b1=1) or (b113c1=1) or (b113d1=1) or (b113e1=1) GBV_emo_e=1.*
- *If (b113a1=0) and (b113b1=0) and (b113c1=0) and (b113d1=0) and (b113e1=0) GBV_emo_e=0.*
- *Variable label GBV_emo_e "GBV_emo_e-Ever experienced emotional violence".*
- *value label GBV_emo_e 1 "Yes" 0 "No".*
- *MISSING VALUE GBV_emo_e (9).*
- *EXECUTE.*
- *IF missing (gbv_emo_e) gbv_emo_e =0.*

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

KIỂM TRA PHÂN BỔ SỐ LIỆU

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Kiểm tra phân bố các biến rời rạc

- Phân bố tần suất (Frequencies)
- Phân bổ phần trăm (Percentages)

- Các biểu đồ phân bố tần suất
  - Biểu đồ cột (Bar charts)
  - Biểu đồ tròn (Pie charts)
  - Histograms

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Mô tả các biến liên tục

- xu hướng tập trung (Central tendency)
  - Mean
  - Median
  - Mode
- Phân bố chuẩn (Normal Distribution)
  - Skewness (tiến gần giá trị 0)
  - Kurtosis (tiến gần giá trị 3)
- Đo lường dàn trải số liệu (Dispersion measures)
  - Range
  - Interquartile range
  - Variance
  - Standard deviation

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Sử dụng Explore để kiểm tra phân bố chuẩn

- Analyze
  - Descriptive statistics
    - explore

- Outlier
- Percentile
- Normal plot with tests
- Stem and leaf

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Làm thế nào để biết một biến liên tục có phân bố chuẩn?

- Sử dụng biểu đồ Histogram với đường cong phân bố chuẩn (Histogram with normal curve)
- Skewness (-: đường cong nghiêng về bên trái, +: đường cong nghiêng về bên phải) và Kurtosis (-: flat, 0: bình thường, +: too pointy)
- Q-Q plot (các số liệu phải nằm trên đường thẳng)
- Kiểm định Kolmogorov-Smirnov

Kiểm tra phân bố chuẩn của số liệu thực hành chủ đề 2

Kiểm tra phân bố chuẩn biến tổng điểm kiến thức về HIV của thanh thiếu niên đường phố (total_kn)

Syntax:

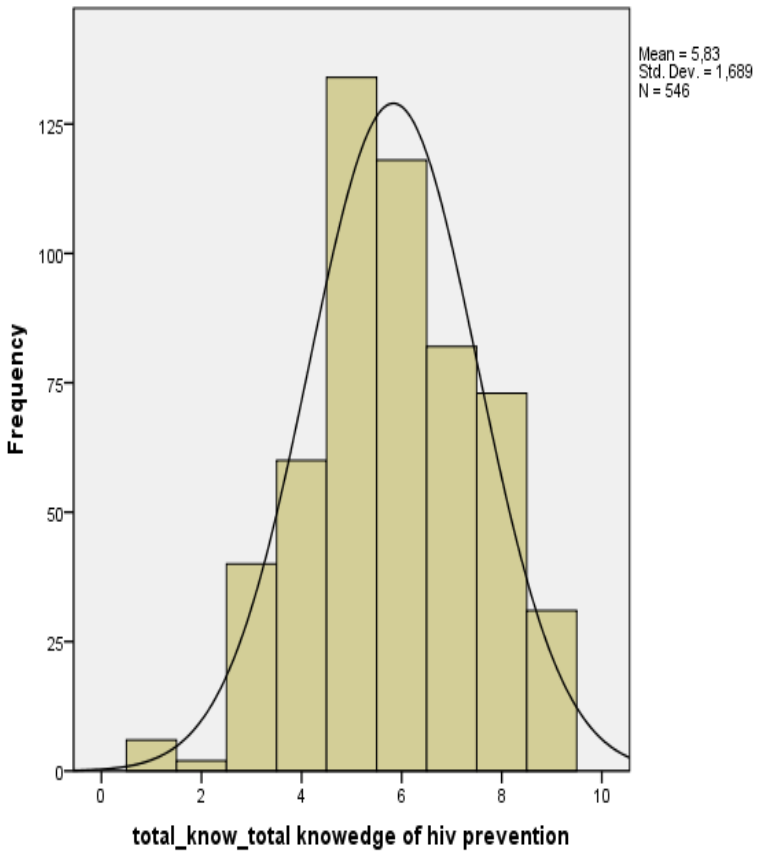
```
EXAMINE VARIABLES=total_kn
  /PLOT BOXPLOT HISTOGRAM
     NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

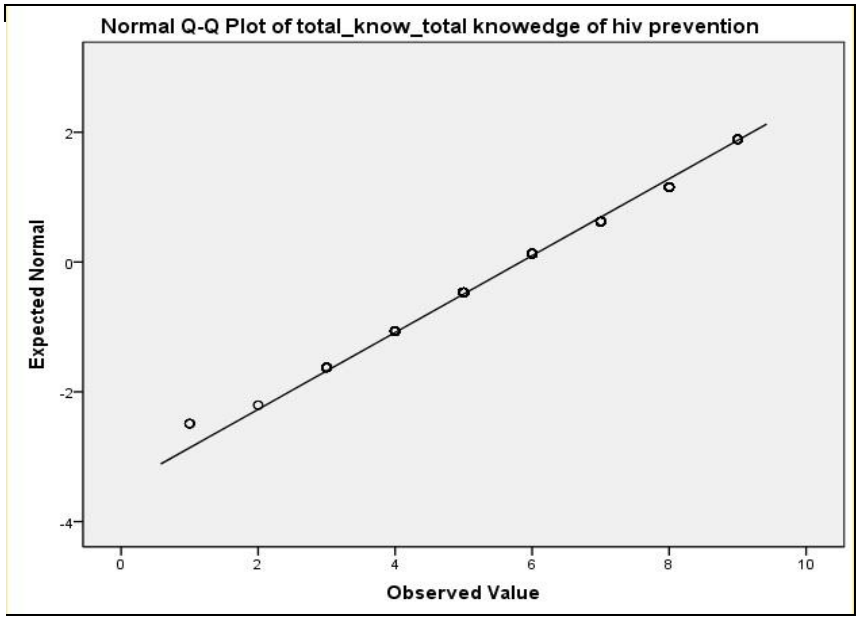
Descriptives

			Statistic	Std. Error
total_know_total knowedge of hiv prevention	Mean		5,83	,072
	95% Confidence Interval for Mean	Lower Bound	5,69	
		Upper Bound	5,98	
	5% Trimmed Mean		5,84	
	Median		6,00	
	Variance		2,851	
	Std. Deviation		1,689	
	Minimum		1	
	Maximum		9	
	Range		8	
	Interquartile Range		2	
	Skewness		-,111	,105
	Kurtosis		-,300	,209

ISMS INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Kiểm tra phân bố chuẩn của số liệu thực hành chủ đề 2 (tiếp)





Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
total_know_total knowedge of hiv prevention	,132	546	,000	,958	546	,000

a. Lilliefors Significance Correction

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Bài viết phân bố chuẩn trên thongke.info

- http://thongke.info.vn/Desktop.aspx/Quan_ly_so_lieu/Phan-bo-chuan-Normal-distribution-trong-Stata/Phan_bo_chuan_Normal_distribution_trong_Stata/

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Xử lý như thế nào nếu biến liên tục không có phân bố chuẩn?

- Natural log
- Square root
- Square
- 1/square root
- Cube

- Sử dụng lệnh tạo biến mới (transform- compute new variables)

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES

Bài tập 2: tạo biến mới

ISMS
INSTITUTE OF SOCIAL AND MEDICAL STUDIES