Lecture 9

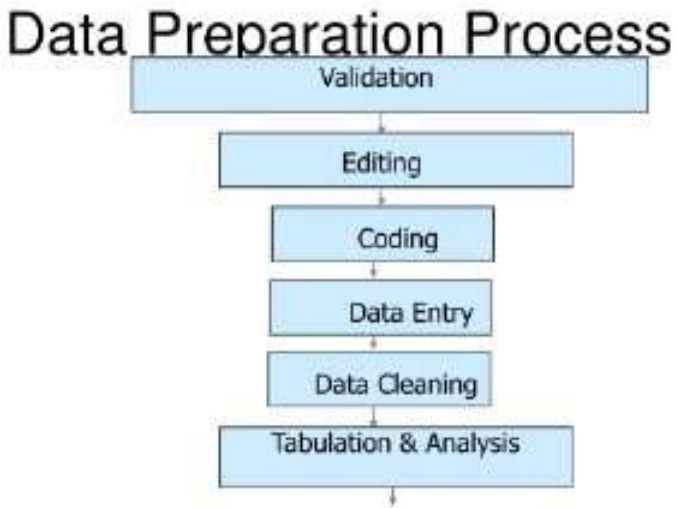
Multivariate Data Analysis

Doszhan R.

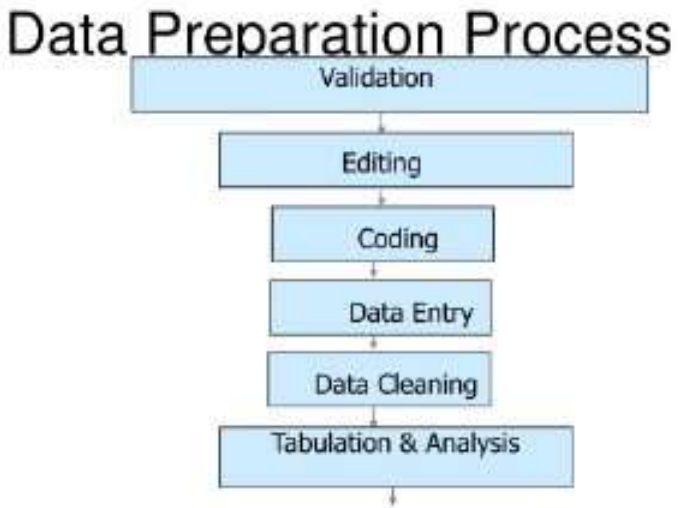
Agenda

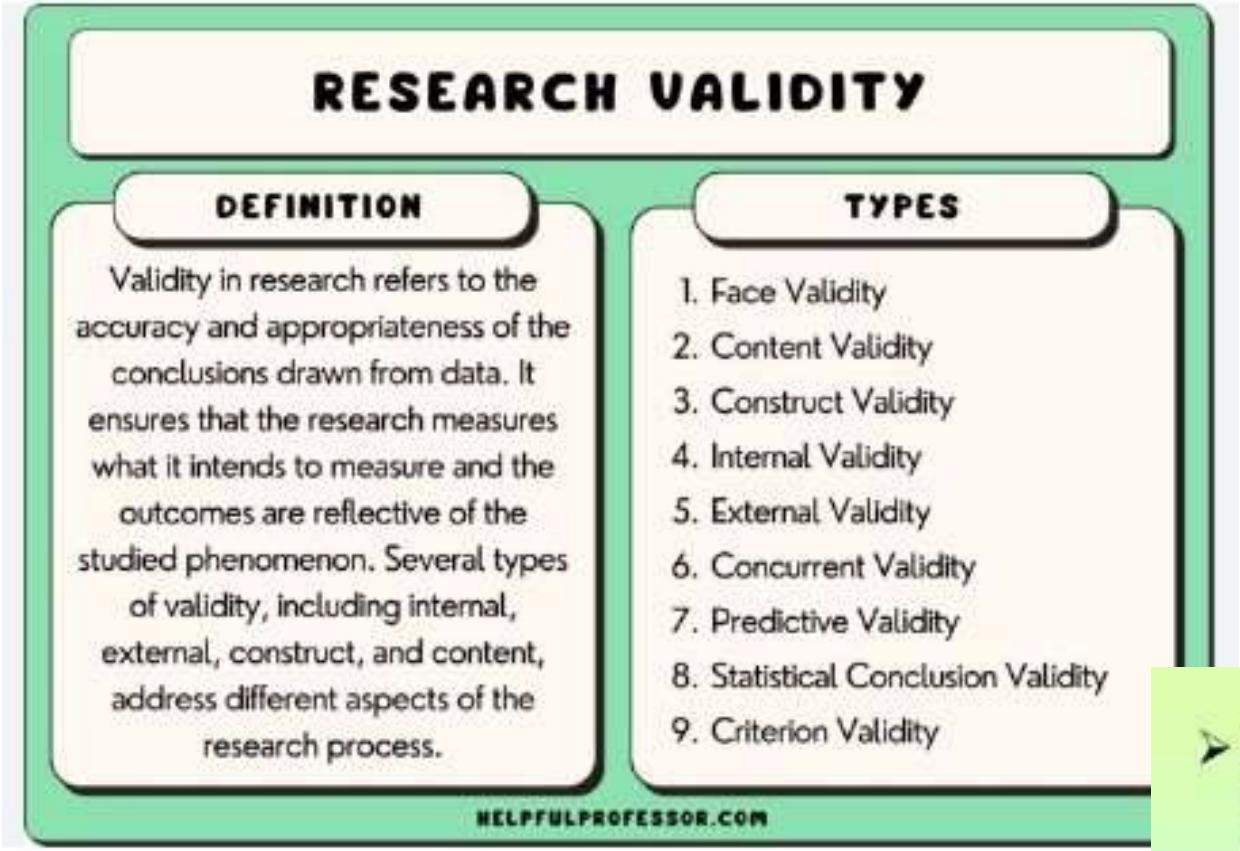
- Data Preparation and Preliminary Analysis
- Validating and Editing
- Coding
- Data Entry
- Data Cleaning
- Tabulation of Survey Results
- Data Mining

Preliminary data preparation techniques



Preliminary data preparation techniques





- Field Editing is a prelimary form of data editing which is undertaken by the field supervisor on the day of the interview with a view to finding omissions, checking the legibility of handwriting, and clarifying responses by respondents that are logically or conceptually inconsistent

- In-House Editing is a form of data editing which is more rigorous than field editing in nature, and which is performed by a centralized office staff

Date: --------
Respondent's telephone number ----------
Respondent's address ----------
Respondent's age:
1. How many chocolates do you eat in a typical week?
A. Less than 5
B. Between 5 and 10
C. Between 11 and 20
D. More than 20
E. Don't know
F. None
(INTERVIEWER- IF RESPONSE IS "A", "E" OR "F", GO TO QUESTION 7)
2. Why do you eat chocolates?
Respondent's answer ----------
3. Which brand of chocolates do you prefer most?
a) Cadbury's
b) Nutrine
c) Nestle
d) Amul
e) Others (specify) ----------
4. When do you like to eat chocolates?
Response ----------
5. Do you prefer chocolates to sweets? (Y/N)
6. Do you have any negative associations with chocolates?
---
7. What is your age group?
a) Under 10
b) Between 10 and 20
c) Between 21 and 30
d) Above 30
e) Refused to answer, no answer or don't know
May I know your name? My office calls about 10% of the people I visit to verify if I have conducted the interviews.
Gave name ----------
Refused to give name: --------
Thank you for your time. Have a good day.

Sample questionnaire consumer survey on chocolate consumption pattern

Treatment of Unsatisfactory Responses:

Assigning missing values,

Returning to the field,

Discarding unsatisfactory responses

Discarding unsatisfactory responses

•Proportion of unsatisfactory respondents is very small compared with the sample size

•Respondents with unsatisfactory responses do not differ from other respondents characteristics

•Unsatisfactory responses for each respondent are proportionately more in each questionnaire

•Responses on key variables are missing

•To reiterate, editing has to be done with patience and care because it is an important step in the processing of questionnaires.

Coding

Coding is the process of assigning numbers or other symbols to answers in order to group the responses into limited categories.

i.e:

‘landlord’ ----- ‘LLD’

‘tenant’ -------‘TNT’

A researcher should follow four rules

1. Appropriate: Categorization should help to validate the hypotheses of the research study.
2. Exhaustive: When multiple-choice questions are used, an adequate list of alternatives should be provided to tap the full range of information from respon- dents.
3. Mutually Exclusive: Complying with this rule requires that a specific alternative is placed in one and only one cell of a category set.
4. Derived: Every class in the category set is defined in terms of one concept.

Code book

Question number	Variable number	Code description	Variable name
1	1	Number of chocolates 1 = Less than 5 2 = Between 5 and 10 3 = Between 10 and 20 4 = Above 20 5 = Don't know 9 = Missing	NO-OF-CHOCOLATES
2	2	Reason(s) 0 = Not mentioned 1 = Mentioned	REASON
2	3	Taste Soft Size Low price Sweet smell Others	TASTE SOFT SIZE COST SMELL OTHERS
3	4	Brand 1 = Cadbury's 2 = Nutrine 3 = Nestle 4 = Amul 5 = others 9 = missing	BRAND
		0 = not mentioned 1 = mentioned	
4	6	Festival Marriage Friends Time pass Happy occasion Bought new goods Gift Others	FESTIVAL MARRIAGE FRIENDS TIMEPASS HAPPY-OCC BOUGHT-NEW GIFT OTHERS
5	7	Preference 1 = yes 0 = no	PREF
6	8	Negative Association 0 = mentioned 1 = not mentioned	ASSOCIATION
6	9	Tooth decay Worms Expiry date Heart disease Others	TOOTH DECAY WORMS EXPIRY-DATE HEART DISEASE OTHERS
7	10	Age group 1 = below 10 2 = between 10 and 20 3 = between 21 and 30 4 = above 30 9 = missing	AGE
		Name	NAME

Example of coding for an open-ended question

Categories
1.Taste
.....
2.Soft
.....
3.Size
.....
4.Low price
.....
5.Sweet smell
.....
6.Others
.....

Table 6.1 Handling DK responses

Question: How many chocolates do you eat in a typical week?
Answers:

Age	Less than 5 (%)	Above 20 (%)	"Don't know" responses (%)
Below 10	29	56	21
10–20	44	24	55
21–30	21	15	17
Above 30	6	5	7
Total n	312 (100)	142 (100)	46 (100)

The ways researchers enter the data

Optical Scanning

Barcode Reader

Voice Recognition

Table 6.2 One-way frequency

Q-3. Which brand of chocolates do you prefer most?	
Brand	Total 500 (100 %)
Cadbury's	225 (45 %)
Nestle	155 (31 %)
Amul	83 (16.6 %)
Nutrine	31 (6.2 %)
Do not know/other	6 (1.2 %)

Table 6.3 Simple cross tabulation

	Ages of respondents				Total
Brand	<10	10–20	21–30	More than 30	
Cadbury's	93	73	39	20	225 (45 %)
Nestle	69	44	22	20	155 (31 %)
Amul	39	20	12	12	83 (16.6 %)
Nutrine	11	9	6	5	31 (6.2 %)
Others	3	3	0	0	6 (1.2 %)
TOTAL	215 (43 %)	149 (29.8 %)	79 (15.8 %)	57 (11.4 %)	500(100 %)

Table 6.4 Cross tabulation

Brand	Number of chocolates	Ages of respondents				Subtotal	Total
		<10	10–20	21–30	More than 30		
Cadbury's	<5	42	34	20	11	107	**225 (45 %)**
	5–10	31	21	12	5	69	
	10–20	15	12	5	2	34	
	More than 20	5	6	2	2	15	
Nestle	<5	36	26	13	12	87	**155 (31 %)**
	5–10	21	12	6	5	44	
	10–20	8	4	2	2	16	
	More than 20	4	2	1	1	8	
Amul	<5	21	11	6	5	43	**83 (16.6 %)**
	5–10	12	5	3	4	24	
	10–20	4	3	2	2	11	
	More than 20	2	1	1	1	5	
Nutrine	<5	5	4	2	3	14	**31 (6.2 %)**
	5–10	3	3	2	1	9	
	10–20	2	1	1	1	5	
	More than 20	1	1	1	0	3	
Others	<5	2	1	0	0	3	**6 (1.2 %)**
	5–10	1	2	0	0	3	
	10–20	0	0	0	0	0	
	More than 20	0	0	0	0	0	
Total		**215 (43 %)**	**149 (29.8 %)**	**79 (15.8 %)**	**57 (11.4 %)**		**500 100 %)**


DATA MINING
Data sets
Pre-processing
Classification
Database
Statistics
Analytics
Evaluation


Healthcare
Market Basket Analysis
Fraud Detection
Lie Detection
Data Mining Applications
CRM
Financial Banking
Manufacturing Engineering
Education