A guide to:
Data collection

Key concepts & study plan

Experimental design

Data collection & processing

Model specification & estimation

Interpretation & application

A guide to: Data collection

Steps in data collection

1. SP and/or RP
2. Sampling strategy and sample size
3. Design of data collection process
4. Ethics approval and consent
5. Pre-testing and pilot
6. Main data collection
7. Database creation

Each study is different, some steps may not apply or in a different order

Step 1 – SP and/or RP

Key concepts
& study plan

Experimental
design

Data collection
& processing

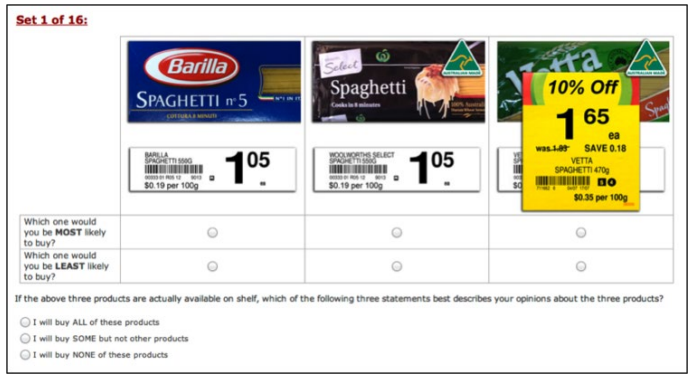
Model specification
& estimation

Interpretation
& application

Step 1 – SP and/or RP

Stated or revealed preference data

- Stated preference data
  - Hypothetical choice context
  - Hypothetical alternatives
  - Hypothetical attributes and levels
  - Stated choices



Study objectives (existing vs new options)
Stakeholders (budget and potential risks)

- Revealed preference data
  - Real-world choice context
  - Real-world alternatives
  - Real-world attributes and levels
  - Revealed choices

Step 2 – Sampling strategy and sample size

Key concepts & study plan

Experimental design

Data collection & processing

Model specification & estimation

Interpretation & application

Step 2 – Sampling strategy and sample size

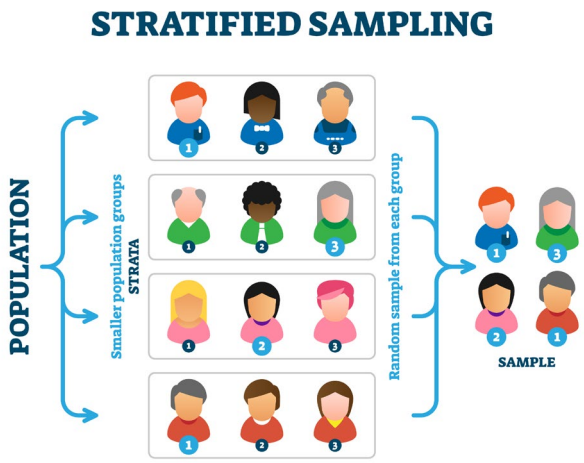
Sample composition

- Target population depends on study objectives
- Collecting data outside target population
  - Is a waste of resources
  - Could bias outcomes
- Efficient sample selection strategies exist

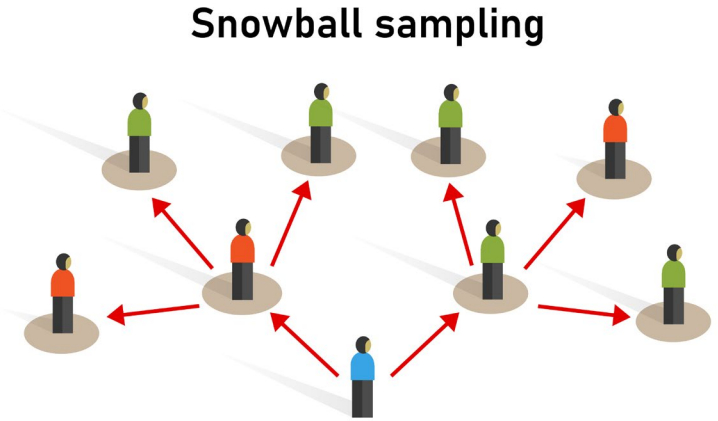
Step 2 – Sampling strategy and sample size

Sampling strategies

- Random sampling is desirable but noisy
- Stratified sampling controls sample characteristics
  - Less noisy
  - Knowledge on population
  - Corrections might be needed



- Snowball sampling



Step 2 – Sampling strategy and sample size

Stratified sampling

Sampling based on respondent features

- Understand preference differences between sub-populations
- Stratified sampling fixes proportion of each sub-population
- Random sampling requires very large sample when one sub-population is small
- Oversampling of small sub-population
  - Improves power of analysis
  - Average sample ≠ average population
  - Corrections are needed (interactions or weighting)

Step 2 – Sampling strategy and sample size

Outcome-based stratified sampling

Outcome dependent sampling

- Sampling conditional on choice

=

- Sampling based on preferences

- Can be efficient
- Requires weighting schemes in analysis

Step 2 – Sampling strategy and sample size

Sample size

- Depends on research question and effect size
- Efficiency of the design
- Desired power of analysis
- Heuristics exist

Step 3 – Design of data collection process

Key concepts & study plan

Experimental design

Data collection & processing

Model specification & estimation

Interpretation & application

Step 3 – Design of data collection process

Data acquisition

- Actual behavior
  - Primary or secondary data collection
  - Choice set information
  - Decision maker features
- Stated preferences
  - Sample acquisition
  - Survey design (next on agenda)
  - Survey implementation
- Data storage and ownership
- Privacy

Step 4 – Ethics approval and consent

Key concepts
& study plan

Experimental
design

**Data collection
& processing**

Model specification
& estimation

Interpretation
& application

Step 4 – Ethics approval and consent

Agreements

Stakeholder and participant agreement

- Ethics and/or Internal Review Board approval
  - Varies across disciplines
  - Required by (some) institutions
  - Required by (some) journals

- Consent
  - Informs participants about the study and how the data will be used
  - Formal decision to participate and share data
  - Record keeping

Step 4 – Ethics approval and consent

Example consent forms for online low risk surveys

- Consent and participant information requirements may vary across countries, institutions, studies

Consent Form

Introduction
This is an academic study conducted by researchers at Erasmus University Rotterdam. Your participation in this study is voluntary and you may withdraw from the study at any time. The study is described as below. Participating in the study might not benefit you, but we might learn things that will benefit others.

Purpose of the study
The purpose of this study is to evaluate how preferences regarding (future) pension income can be most accurately measured. This information helps pension funds in determining the appropriate investment strategy that best matches the risk-return preferences of their participants.

What you will be asked to do
You will be asked to make multiple hypothetical decisions regarding what pension you would prefer in terms of risks and returns. This survey will take about 15-20 minutes.

Possible risks and discomforts
Minimal risks, which may include fatigue or eye strain, may occur. The risks or discomforts anticipated are not beyond what you may expect to experience in your daily lives.

Confidentiality & Anonymity
Your responses are anonymous and thus results will be reported with no reference to you specifically. No information that could allow the researcher to identify you personally will be collected. Therefore, your responses cannot be linked to your identity.
Once the study has been completed, all data files will be safely stored on a server at Erasmus University Rotterdam in the Netherlands.

Questions or Concerns
If you have any questions regarding the study, please contact the study administrator at donkers@ese.eur.nl.

Consent
Please click on the **I AGREE** button below, if you have understood to your satisfaction the information regarding participation in the research project, that you are aware that all records are entirely confidential and that you may discontinue participation at any point in the study, and that you agree to participate.

Consent

You are consenting to take part in this research as follows:

I have downloaded, read and understood the information provided in the Participant Information Statement.

I agree to participate in this survey, realising that I can withdraw at any time while completing the survey questions without adverse consequences. I understand that once I have completed the survey, I cannot withdraw my consent as the survey is anonymous.

I agree that research data collected for the study may be published or may be provided to other researchers in a form that does not identify me in any way.

Select only one answer

- I agree and consent to participate
- I do not agree and/or do not consent to participate

Previous Question | Next Question

Step 5 – Pre-testing and pilot

Key concepts & study plan

Experimental design

Data collection & processing

Model specification & estimation

Interpretation & application

Step 5 – Pre-testing and pilot

Pre-testing

Qualitative check on survey quality

- Understanding / clarity
- Feedback
- Think-aloud study

Step 5 – Pre-testing and pilot

Pilot study

- Small scale test
  - ≈10% of sample
- Check on understanding
- Choice/response patterns
- Check estimation routine
- Advanced: update your design

Step 6 – Main data collection

Step 6 – Main data collection

Main data collection

- It is out of your hands now
  - Preparation is everything

- Timing
- Quota

Step 7 – Database creation

Key concepts
& study plan

Experimental
design

**Data collection
& processing**

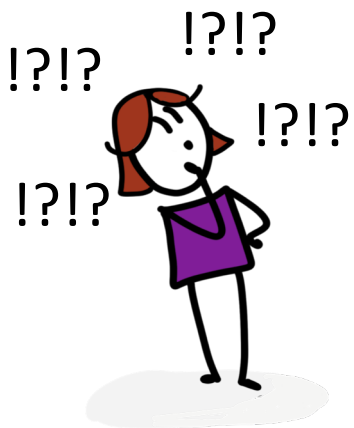
Model specification
& estimation

Interpretation
& application

Step 7 – Database creation

Data cleaning

- Data cleaning is only 2nd best option
- Invest in data quality from the start
  - Incentive compatibility
- Survey design and simplicity (instead of complexity) are key
  - Pre-test with think aloud



Step 7 – Database creation

Data cleaning

- Remove observations only with care to avoid wasting data and/or bias results

- Bad data examples
  - Straightliners (e.g., always choosing left option)
  - Speeders (e.g., choosing within seconds)
  - Always choose cheapest (or is this true preference)
  - Incomplete responses
  - Unrealistic responses
  - Inconsistent responses
  - Nonsensical responses

Step 7 – Database creation

Example survey data

- See `example dataset.xlsx`

In what domain(s) do you work?

Select all that apply

- [ ] Health
- [ ] Transport
- [ ] Environmental economics
- [ ] Marketing
- [ ] Food
- [ ] Other

	Pet A	Pet B
Type of animal	Porcupine 	Tarantula 
Envy factor	We all got one during Covid	Talk of the town
Size of enclosure	Large: no more guest room	Small: fits on the kitchen counter
Risk of accidents	Medium: 30% risk of small bruises and scratches	High: 80% risk of getting hit, stung or bitten
Monthly expenses	100 €	100 €
Which would you choose?	○	○

Step 7 – Database creation

Data setup

- Long format – each row contains data of a single alternative in a choice task
- Used in Nlogit

choice task

respondent

RID	duration	health	transport	environment	marketing	food	other	petowner	design_row	choicetask	csetsize	choice	animal	envy	size	risk	cost
6	193.464	1	1	1	2	2	1	2	7	1	2	0	5	3	2	1	100
6	193.464	1	1	1	2	2	1	2	7	1	2	1	1	2	1	1	10
6	193.464	1	1	1	2	2	1	2	17	2	2	1	5	3	2	1	10
6	193.464	1	1	1	2	2	1	2	17	2	2	0	2	1	2	3	100
6	193.464	1	1	1	2	2	1	2	6	3	2	1	5	2	1	2	10
6	193.464	1	1	1	2	2	1	2	6	3	2	0	4	1	3	2	50
6	193.464	1	1	1	2	2	1	2	16	4	2	1	2	2	2	3	100
6	193.464	1	1	1	2	2	1	2	16	4	2	0	5	1	2	3	100
6	193.464	1	1	1	2	2	1	2	15	5	2	0	3	2	3	3	500
6	193.464	1	1	1	2	2	1	2	15	5	2	1	5	3	3	2	50
6	193.464	1	1	1	2	2	1	2	13	6	2	1	5	2	3	1	10
6	193.464	1	1	1	2	2	1	2	13	6	2	0	2	2	2	3	100
6	193.464	1	1	1	2	2	1	2	14	7	2	1	4	3	1	3	10
6	193.464	1	1	1	2	2	1	2	14	7	2	0	3	2	1	2	100
6	193.464	1	1	1	2	2	1	2	8	8	2	0	5	2	2	1	500
6	193.464	1	1	1	2	2	1	2	8	8	2	1	3	2	2	2	100
9	861.232	1	1	2	1	1	1	2	12	1	2	0	3	2	2	1	50
9	861.232	1	1	2	1	1	1	2	12	1	2	1	5	3	1	3	100
9	861.232	1	1	2	1	1	1	2	4	2	2	1	1	2	2	3	10
9	861.232	1	1	2	1	1	1	2	4	2	2	0	2	2	1	2	10
9	861.232	1	1	2	1	1	1	2	20	3	2	1	4	2	3	2	100
9	861.232	1	1	2	1	1	1	2	20	3	2	0	5	1	3	1	10
9	861.232	1	1	2	1	1	1	2	27	4	2	0	5	2	1	2	50
9	861.232	1	1	2	1	1	1	2	27	4	2	1	2	2	3	3	10

data (long format) | data (wide format) | dictionary

Step 7 – Database creation

Data setup

- Wide format – each row contains data of all alternatives in a choice task
- Used in Apollo and Biogeme

choice task

respondent

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	RID	duration	health	transport	environment	marketing	food	other	petowner	design_row	choicetask	csetsize	choice	A_animal	A_envy	A_size	A_risk	A_cost	B_animal	B_envy	B_size	B_risk	B_cost
2	6	193.464	1	1	1	2	2	1	2	7	1	2	2	5	3	2	1	100	1	2	1	1	10
3	6	193.464	1	1	1	2	2	1	2	17	2	2	1	5	3	2	1	10	2	1	2	3	100
4	6	193.464	1	1	1	2	2	1	2	6	3	2	1	5	2	1	2	10	4	1	3	2	50
5	6	193.464	1	1	1	2	2	1	2	16	4	2	1	2	2	2	3	100	5	1	2	3	100
6	6	193.464	1	1	1	2	2	1	2	15	5	2	2	3	2	3	3	500	5	3	3	2	50
7	6	193.464	1	1	1	2	2	1	2	13	6	2	1	5	2	3	1	10	2	2	2	3	100
8	6	193.464	1	1	1	2	2	1	2	14	7	2	1	4	3	1	3	10	3	2	1	2	100
9	6	193.464	1	1	1	2	2	1	2	8	8	2	2	5	2	2	1	500	3	2	2	2	100
10	9	861.232	1	1	2	1	1	1	2	12	1	2	2	3	2	2	1	50	5	3	1	3	100
11	9	861.232	1	1	2	1	1	1	2	4	2	2	1	1	2	2	3	10	2	2	1	2	10
12	9	861.232	1	1	2	1	1	1	2	20	3	2	1	4	2	3	2	100	5	1	3	1	10
13	9	861.232	1	1	2	1	1	1	2	27	4	2	2	5	2	1	2	50	2	2	3	3	10
14	9	861.232	1	1	2	1	1	1	2	18	5	2	1	5	2	3	3	100	2	2	3	2	500
15	9	861.232	1	1	2	1	1	1	2	19	6	2	1	1	2	1	3	500	2	2	1	1	500
16	9	861.232	1	1	2	1	1	1	2	26	7	2	1	3	1	2	3	100	2	2	2	1	50
17	9	861.232	1	1	2	1	1	1	2	5	8	2	2	2	2	2	2	500	1	1	2	2	50
18	10	725.446	1	1	1	1	1	2	2	9	1	2	1	3	1	3	2	50	1	2	1	2	10
19	10	725.446	1	1	1	1	1	2	2	2	2	2	2	5	3	1	2	500	3	2	2	2	50
20	10	725.446	1	1	1	1	1	2	2	11	3	2	1	2	1	2	2	500	4	3	2	2	10
21	10	725.446	1	1	1	1	1	2	2	24	4	2	1	2	1	1	1	50	3	1	1	1	50
22	10	725.446	1	1	1	1	1	2	2	22	5	2	2	5	2	3	2	100	2	1	3	2	500
23	10	725.446	1	1	1	1	1	2	2	23	6	2	2	5	3	1	2	500	4	3	1	1	100
24	10	725.446	1	1	1	1	1	2	2	3	7	2	2	5	1	2	3	50	3	1	3	1	100
25	10	725.446	1	1	1	1	1	2	2	21	8	2	2	5	2	2	1	100	3	2	3	2	10

data (long format) | data (wide format) | dictionary

Step 7 – Database creation

Data dictionary

- Explains each variable in the data set
  - Descriptions for categorical variables
  - Units for numerical variables

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		**COVARIATES**					**CHOICE EXPERIMENT**									
2																
3		*variable*	*description*				*variable*	*description*								
4																
5		RID	Unique respondent identifier				design_row	Row number in experimental design								
6		duration	Time that respondent spent in survey		Seconds		choicetask	Choicetask number								
7		health	Health	1	Selected		csetsize	Choice set size								
8				2	Not selected		choice	Which would you choose?	1	Pet A						
9		transport	Transport	1	Selected				2	Pet B						
10				2	Not selected		animal	Type of animal	1	Kangaroo <img src="kangaroo.jpg">						
11		environment	Environmental economics	1	Selected				2	Monkey <img src="monkey.jpg">						
12				2	Not selected				3	Tortoise <img src="tortoise.jpg">						
13		marketing	Marketing	1	Selected				4	Tarantula <img src="tarantula.jpg">						
14				2	Not selected				5	Porcupine <img src="porcupine.jpg">						
15		food	Food	1	Selected		envy	Envy factor	1	Talk of the town						
16				2	Not selected				2	We all got one during Covid						
17		other	Other	1	Selected				3	Ewww!!						
18				2	Not selected		size	Size of enclosure	1	Very large: takes up 80% of your living room						
19		petowner	First up, are you a pet owner?	1	Yes				2	Large: no more guest room						
20				2	No				3	Small: fits on the kitchen counter						
21							risk	Risk of accidents	1	High: 80% risk of getting hit, stung or bitten						
22									2	Medium: 30% risk of small bruises and scratches						
23									3	Low: will not engage						
24							cost	Monthly expense		Euros						
25																

data (long format) | data (wide format) | dictionary