Seminar in Industrial Economics (EE489)
Methodology

Dr. Wanwiphang Manachotphong

Working with data

1. Define a research question.
2. Choose a methodology
3. Find the data
4. Analyze the data

(Sometimes, the reverse is true. This is not ideal, though)

Types of data

IN TERMS OF STRUCTURE

-Cross-sectional

-Panel (longitudinal)

-Time-series

IN TERMS OF UNIT OF ANALYSIS

- Macro data (usually at the country or regional level)

- Micro data (for example, individual, household, firm, hospital, school, sport team, etc.)

NON-EXPERIMENT VS. EXPERIMENT

TABLE 1.1

A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
⋮	⋮	⋮	⋮	⋮	⋮
525	11.56	16	5	0	1
526	3.50	14	5	1	0

TABLE 1.2

A Data Set on Economic Growth Rates and Country Characteristics

obsno	country	gpcrgdp	govcons60	second60
1	Argentina	0.89	9	32
2	Austria	3.32	16	50
3	Belgium	2.56	13	69
4	Bolivia	1.24	18	12
⋮	⋮	⋮	⋮	⋮
61	Zimbabwe	2.30	17	6

TABLE 1.4

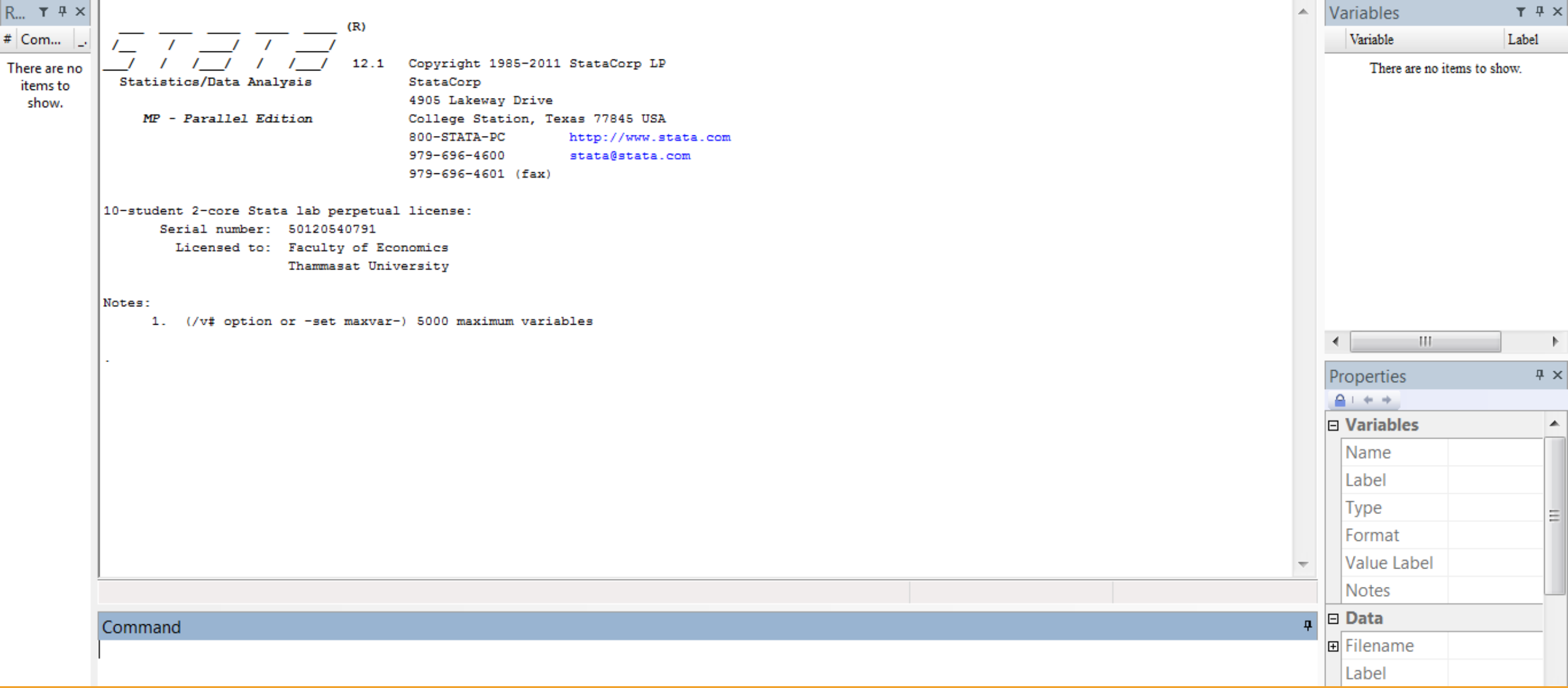
Pooled Cross Sections: Two Years of Housing Prices

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
520	1995	57200	16	1100	2	1.5

TABLE 1.5

A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
⋮	⋮	⋮	⋮	⋮	⋮	⋮
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493



Data and Applying Econometrics

Microeconometrics

- Econometric techniques which can be used to analyze micro-level data.
- Problems usually found in micro-level data:
  - *Simultaneity problem*
  - *Unobserved heterogeneity problem*
  - *Individual's complexity of choices*
  - *Individual's heterogeneity*

Useful textbooks for microeconometrics

Microeconometrics: Methods and Applications by Cameron and Trivedi

Microeconometrics Using Stata by Cameron and Trivedi

Econometrics of Cross Section and Panel Data by Wooldridge, J. M. (or Introductory Econometrics)

Applied Time Series Econometrics by W. Enders

Discrete Choice Methods with Simulation by Kenneth E. Train

Econometric Methods with Applications in Business and Economics by C. Heij and others

Guides to Econometrics by Peter Kennedy

Mostly Harmless Econometrics: An empiricist's Companion by Angrist and Pischke

Econometrics for time-series data

- Econometric techniques which can be used to analyze time-series data.
- Problems usually found in time-series data:
  - *Auto-correlation*
  - *Non-stationarity*
  - *Spurious correlation*
  - *Trend, Cycle, Season effects.*

Useful textbooks for time-series econometrics

The Cointegrated VAR Model: Methodology and Applications by Katarina Jeselius

Time Series Analysis by James D. Hamilton

Introduction to Econometrics by J.H.Stock and M.W.Watson

Time Series Analysis and its Applications. With R Examples by Shumway and Stoffer

What kind of data do you expect to use?

Which empirical method(s) do you think you will use?

Today's materials

Microeconometrics

- Basic OLS
- Differences-in-Differences (Treatment effects)
- Discrete-choice methods / Count models

Survey Design

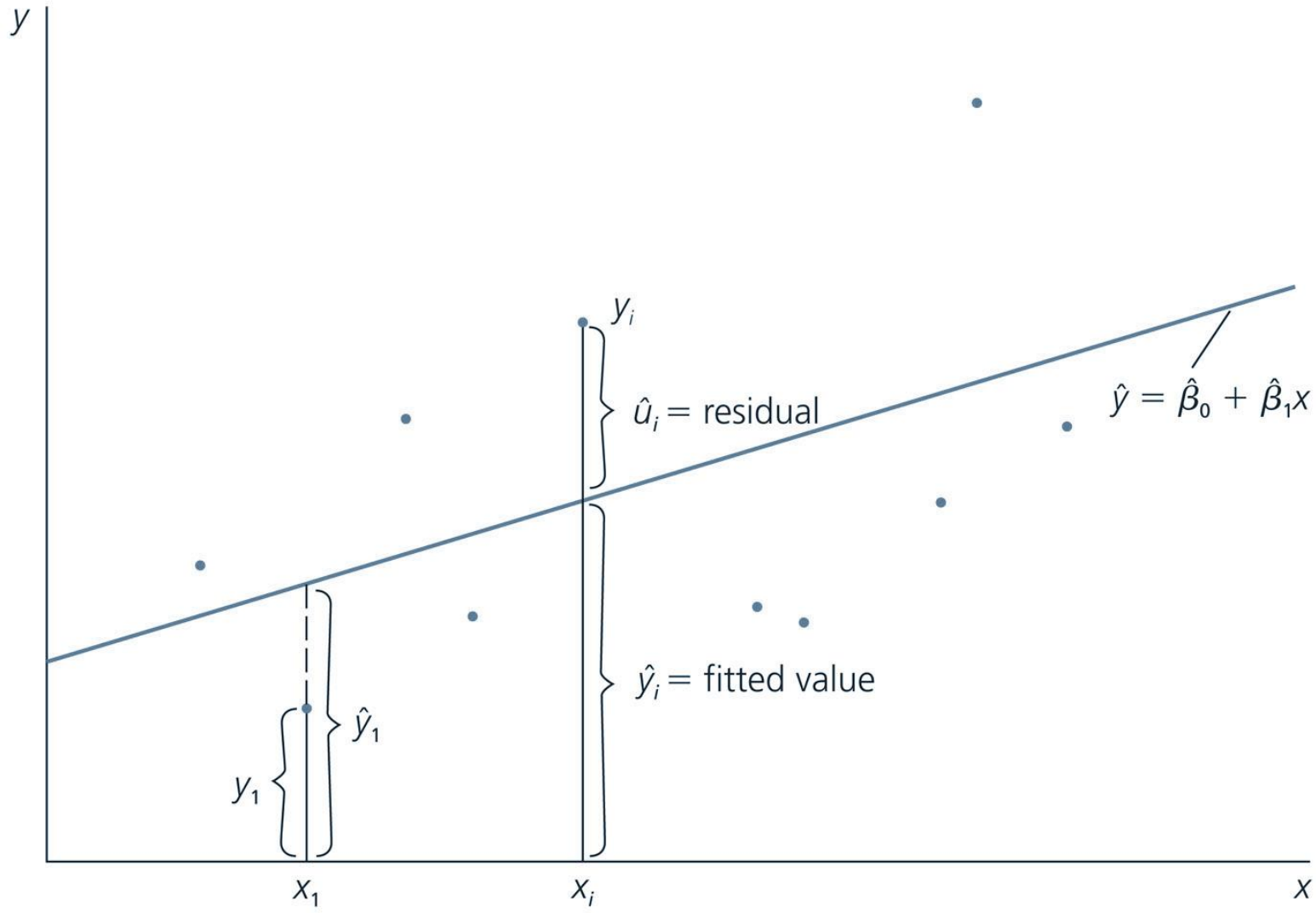
Terminology for Simple Regression

y	x
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor

Basic OLS

FIGURE 2.4

Fitted values and residuals.



Fitted Values and Residuals

TABLE 2.2

Fitted Values and Residuals for the First 15 CEOs

obsno	*roe*	*salary*	*salaryhat*	*uhat*
1	14.1	1095	1224.058	−129.0581
2	10.9	1001	1164.854	−163.8542
3	23.5	1122	1397.969	−275.9692
4	5.9	578	1072.348	−494.3484
5	13.8	1368	1218.508	149.4923
6	20.0	1145	1333.215	−188.2151
7	16.4	1078	1266.611	−188.6108
8	16.3	1094	1264.761	−170.7606
9	10.5	1237	1157.454	79.54626
10	26.3	833	1449.773	−616.7726
11	25.9	567	1442.372	−875.3721
12	26.8	933	1459.023	−526.0231

Basic Issues to consider

1. Heteroskedasticity
2. Omitted Variable Bias
3. Endogeneity Bias (x explains y, while y also explains x)

OLS can easily be done on STATA.

Command: regress y x_1 x_2 x_3 x_4, robust

Differences-in-Differences

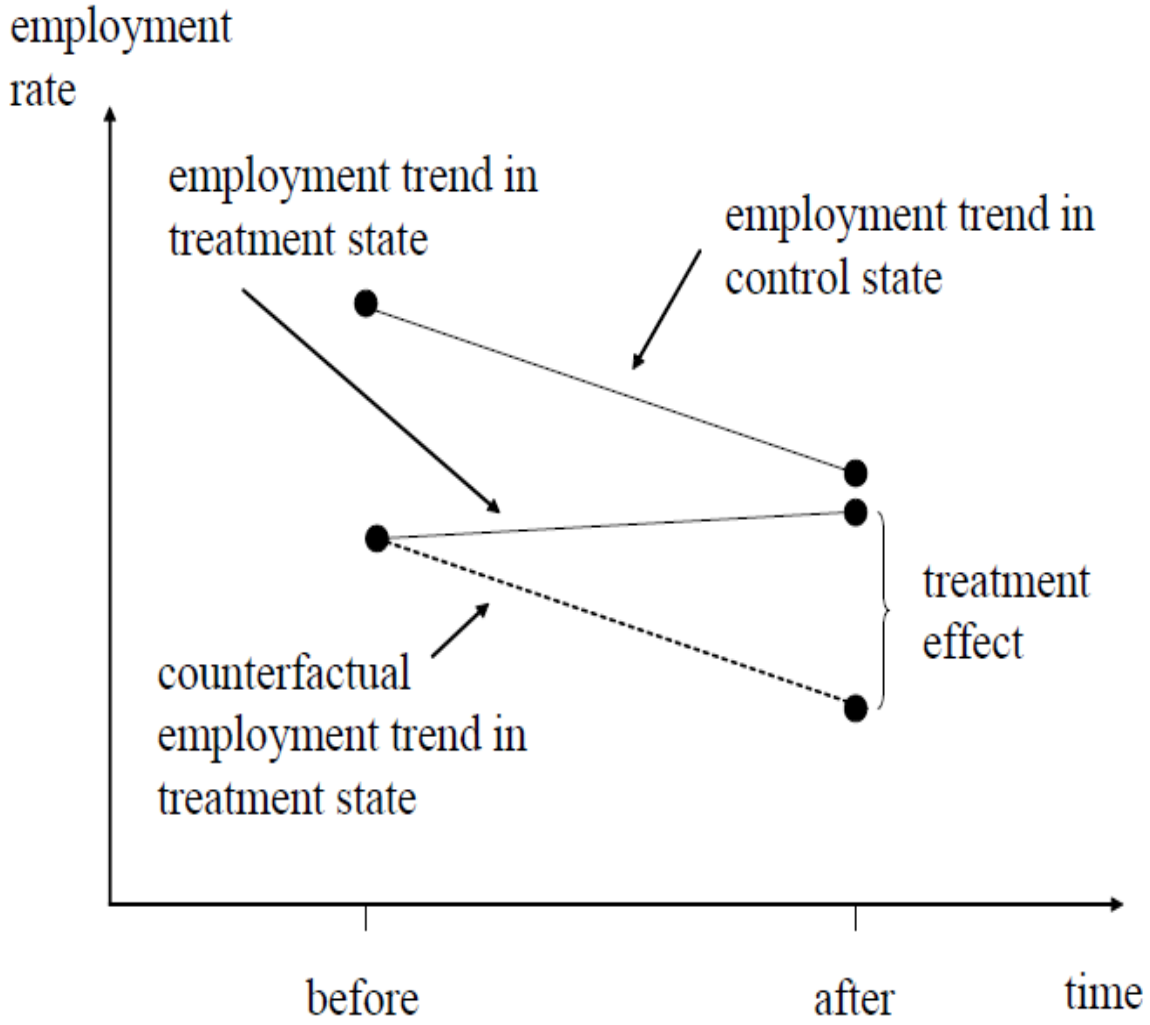
Popular among policy/impact evaluation work.

i.e.

- What is the health impact of Thailand's universal health care policy (30-THB policy)?
- Does an increase in the minimum wage reduce employment?
- Can the rice pledging policy help alleviate farmers' debt problem?

Sometimes we call this the measurement of "treatment effect"

	Control Samples	Treatment Samples
Time = 0	No Policy	No policy
Time = 1	No Policy	With Policy



Source: Angrist and Pischke (2009), p.231

Card and Krueger (2000)

- Measures the impact of an increase in the minimum wage in the US.
- The minimum wage affected Pennsylvania, but not New Jersey.

Regression DD

The specification can turn out to look like a fixed-effect specification with additional period-specific and treatment dummies.

$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ_s \times d_t) + \varepsilon_{ist}$$

where

Y_{ist} is the outcome variable of interest (i.e. employment)

NJ_s is a dummy for each state (here, New Jersey)

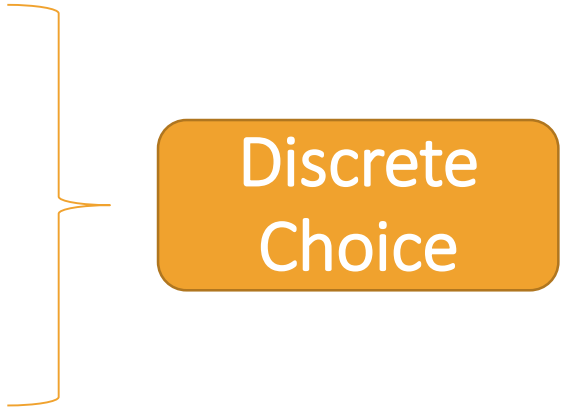
d_t is a dummy for each time period

Thus, **δ measures the treatment effect.**

Discrete Choice methods

Popular discrete-choice methods

- Logit/Probit
- Multinomial Logit
- Ordered Logit
- Nested Logit
- Negative Binomial
- Poisson Regression
- Censored/truncated data



Which model to use depends on the nature of the decision-making process and the distribution of the data.

The Logit/Probit family

- Logit/Probit – for binary choice (yes/no)
- Multinomial Logit – for multiple choice that are not rankable (but has to satisfy the independence of irrelevant alternatives (IIA)
- Ordered Logit – for choices that are rankable
- Nested Logit – for choices that have some nested structures.

Logit & Probit Models

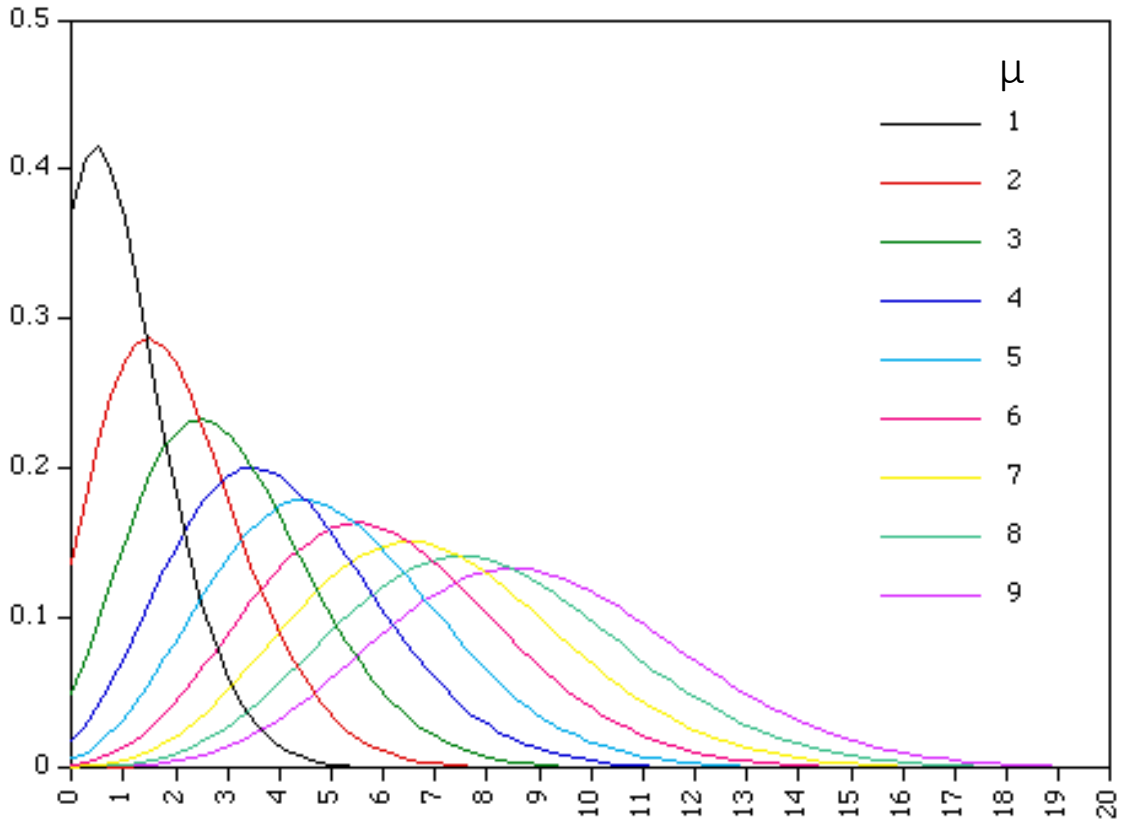
$$P(Y = 1 \mid \mathbf{X}) = G(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k),$$

where $G(\cdot)$ is the cumulative normal distribution function in the probit model. $G(\cdot)$ is the logistic function in the logit model.

Count models

- Sample concentrates on many small count values.
  - doctor visits
  - patent applications

- Popular models
  - Poisson
  - Negative binomials



Deb and Trivedi (2002)

- Data from the RAND Health Insurance Experiment
- Finds the impact of the change of insurance plan on # visits.

Contacts	0	1	2	3	4	5	6	7	8	9	10
Relative Frequency	31.2	18.9	13.8	9.3	6.7	4.8	3.4	2.6	2.0	1.4	1.0
Contacts	11	12	13	14	15	16	…	>21	Max		
Relative Frequency	0.9	0.6	0.5	0.4	0.3	0.3		1.0	77		

Source: Cameron and Trivedi (2005)

Poisson Regression Model

Poisson distribution

$$\Pr[Y = y] = \frac{e^{-\mu}\mu^{y}}{y!},\ y = 0,1,2,....$$

where E(Y) = μ and V(Y) = μ

Poisson MLE finds β that corresponds to the following first order condition. (see more in Cameron and Trivedi (2005))

$$\sum_{i=1}^{N}(y_i - \exp(x_i'\beta))x_i = 0$$

Negative Binomial

- The simple poisson regression requires restrictive assumptions
- Negative binomial is a mixture model which relaxes the assumption that parameter "μ" is fixed.

Poisson pdf.

$$f(y \mid \mu) = \frac{e^{-\mu}\mu^{y}}{y!},\ y = 0,1,2,....$$

The parameter "μ" can become "μ = αv", where α is a deterministic function of X (individual characteristics, explanatory variables). (see more in Cameron and Trivedi (2005))

Sampling & Questionnaire Design

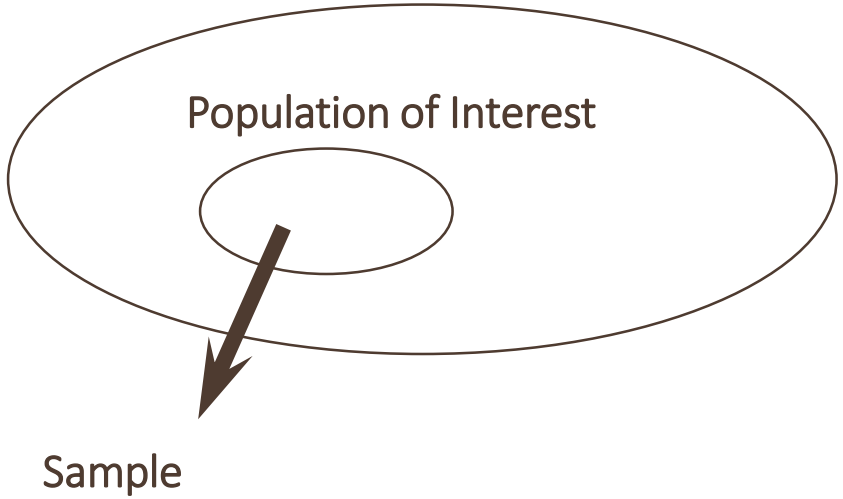
Sampling Design

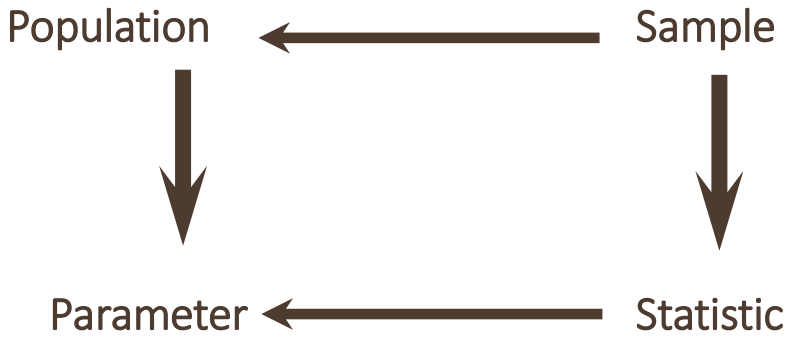
- Obtain information from a subset of a larger group
- The results used to make estimates of the larger group
- Faster and cheaper than asking the entire population

Two keys

1. Selecting the right people
2. Selecting the right number of the right people

Population vs. Sample





We measure the sample using statistics in order to draw inferences about the population and its parameters.

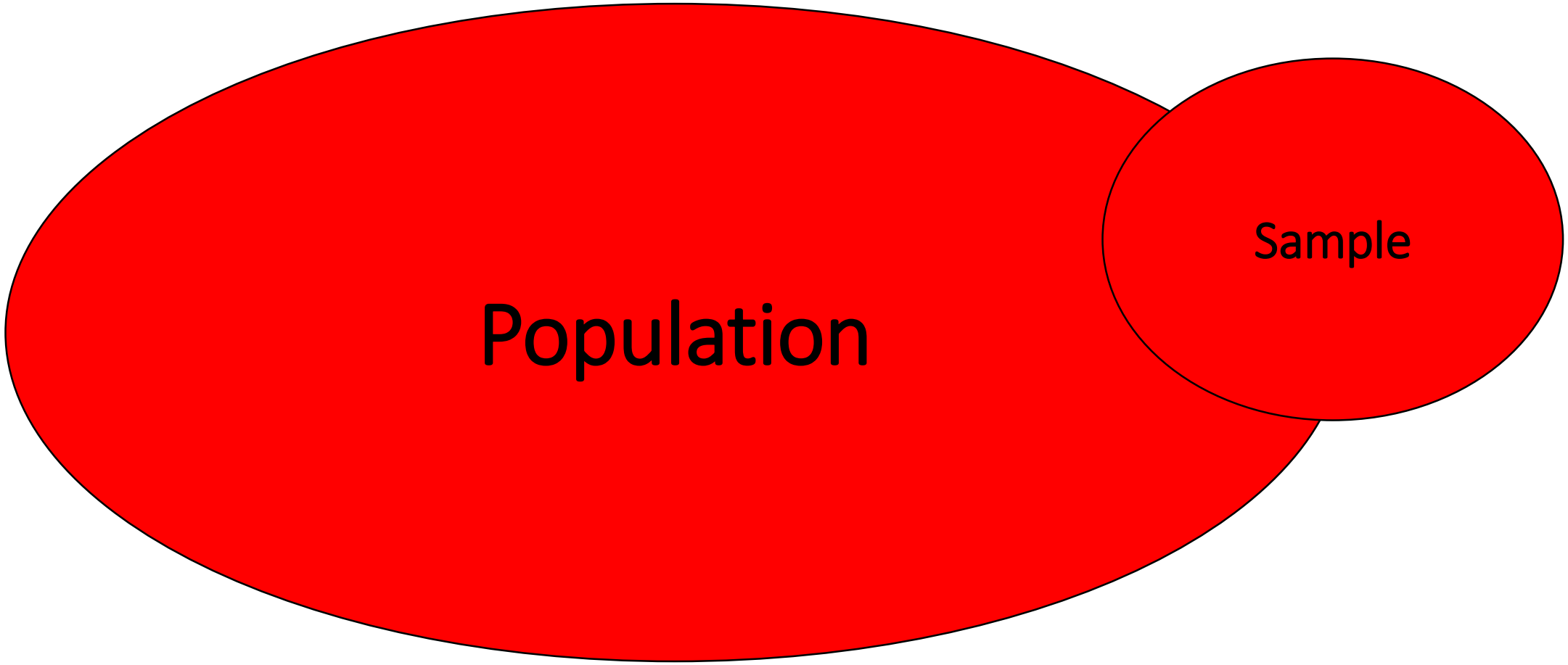
Characteristics of Good Samples

Representative

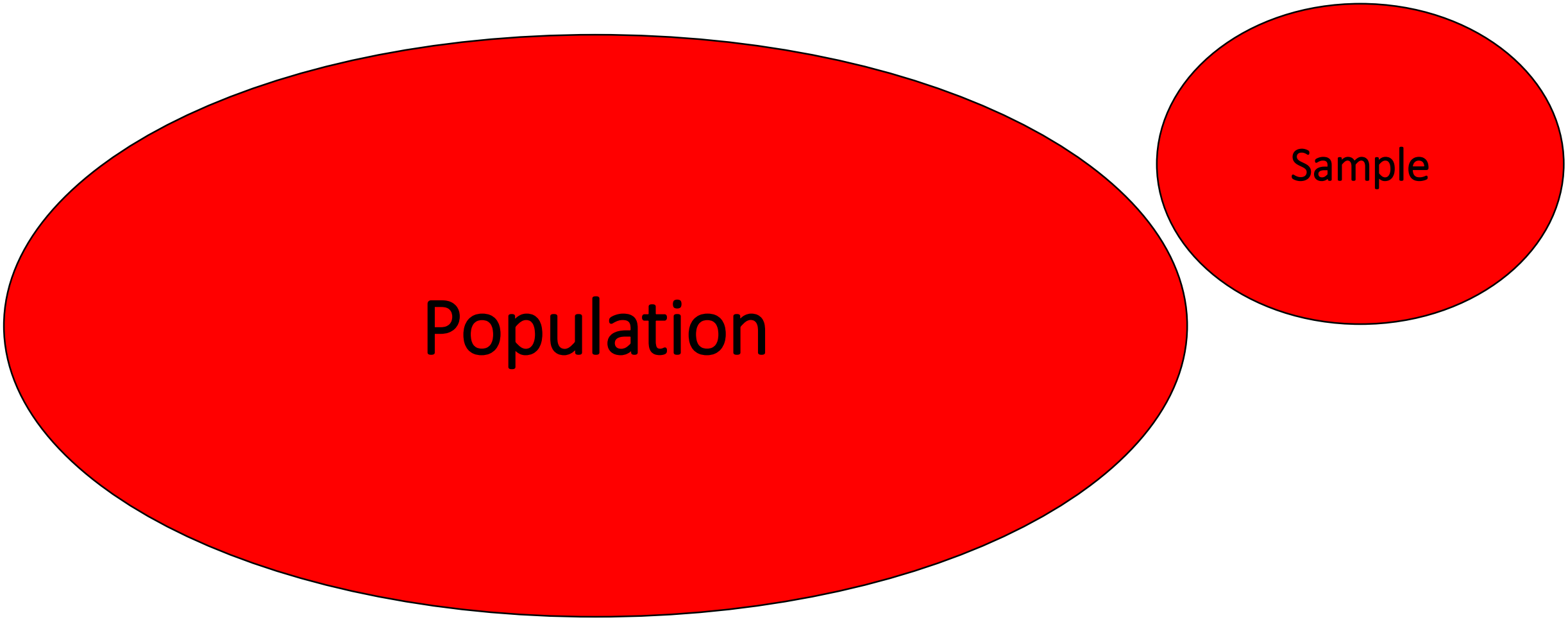
Accessible

Low cost

…This (bad)…



...or this (VERY bad)...



Terminology

- *Population*
  The entire group of people of interest from whom the researcher needs to obtain information.
- *Element (sampling unit)*
  one unit from a population
- *Sampling*
  The selection of a subset of the population
- *Sampling Frame*
  Listing of population from which a sample is chosen
- *Census*
  A polling of the entire population
- *Survey*
  A polling of the sample

Terminology

- *Parameter*

  The variable of interest

- *Statistic*

  The information obtained from the sample about the parameter

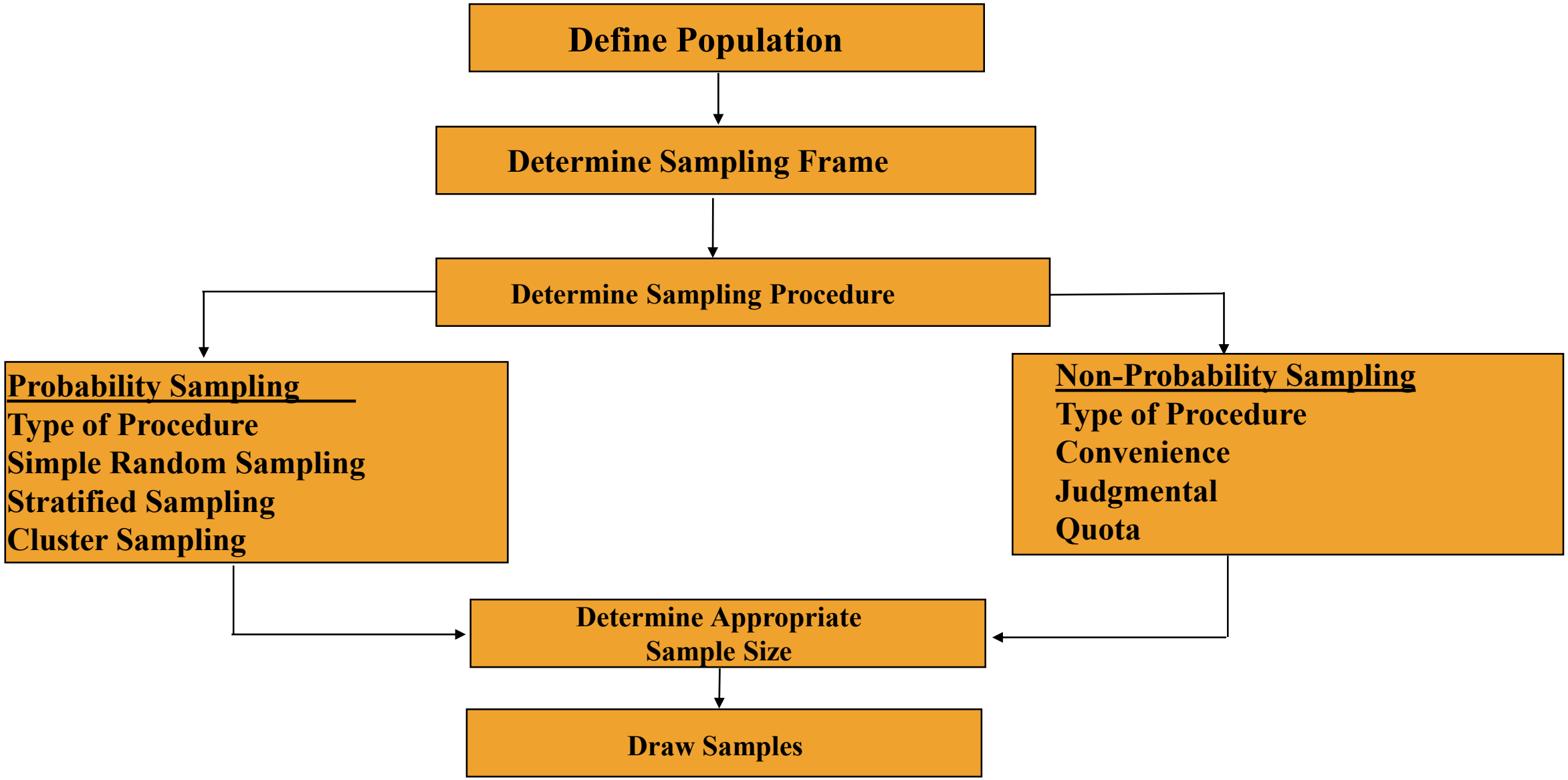
- *Goal*

  To be able to make inferences about the population parameter from knowledge of the relevant statistic - to draw general conclusions about the entire body of units

- *Critical Assumption*

  The sample chosen is representative of the population

Sampling Design Process



1.Define the target population

- "Ideally, who do you want to survey?"

  i.e. those who have the information sought What are their characteristics. Who should be excluded?

- It involves
  - defining population units
  - setting population boundaries
  - Screening (e.g. security questions, product use)

2. Determine the Sampling Frame

Select “*sample units*”

- Individuals
- Household
- Streets
- Telephone numbers
- Companies

3. Selecting a Sampling Design

PROBABILITY SAMPLING

Equal chance of being included in the sample (random)

- simple random sampling
- systematic sampling
- stratified sampling
- cluster sampling

NON-PROBABILITY SAMPLING

Unequal chance of being included in the sample (non-random)

- convenience sampling
- judgment sampling
- snowball sampling
- quota sampling

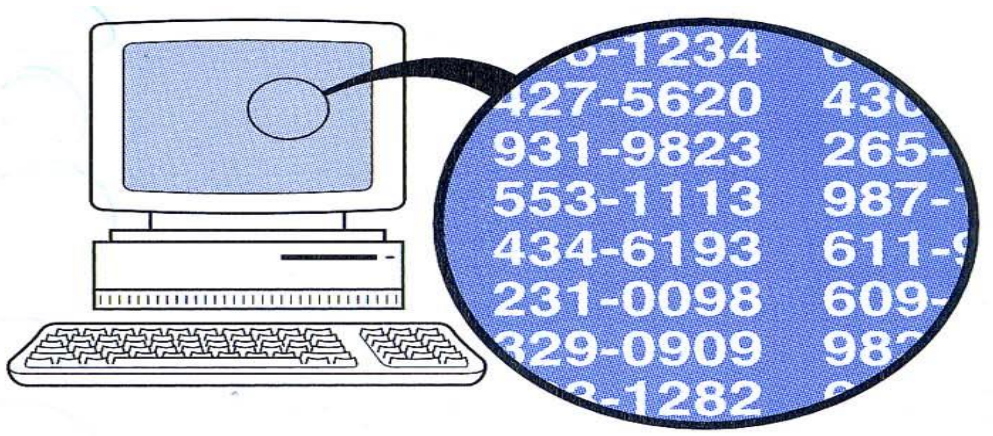
• Simple Random Sampling (SRS)

Population members are selected directly from the sampling frame

Equal probability of selection for every member (sample size/population size)

400/10,000 = .04

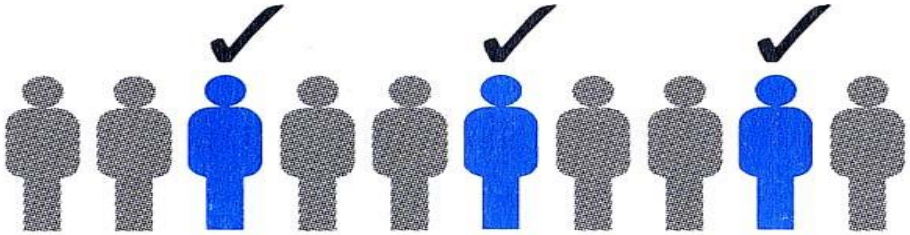
Use random number table or random number generator



• Systematic Sampling

Order all units in the sampling frame based on some variable and number them from 1 to *N*

Choose a random starting place from 1 to *N* and then sample every k units after that



• Systematic Sampling

N = 100

want n = 20

N/n = 5

select a random number from 1-5: chose 4

start with #4 and take every 5th unit

1	26	51	76
2	27	52	77
3	28	53	78
4	29	54	79
5	30	55	80
6	31	56	81
7	32	57	82
8	33	58	83
9	34	59	84
10	35	60	85
11	36	61	86
12	37	62	87
13	38	63	88
14	39	64	89
15	40	65	90
16	41	66	91
17	42	67	92
18	43	68	93
19	44	69	94
20	45	70	95
21	46	71	96
22	47	72	97
23	48	73	98
24	49	74	99
25	50	75	100

• Stratified Sampling (I)

The chosen sample is forced to contain units from each of the segments, or strata, of the population

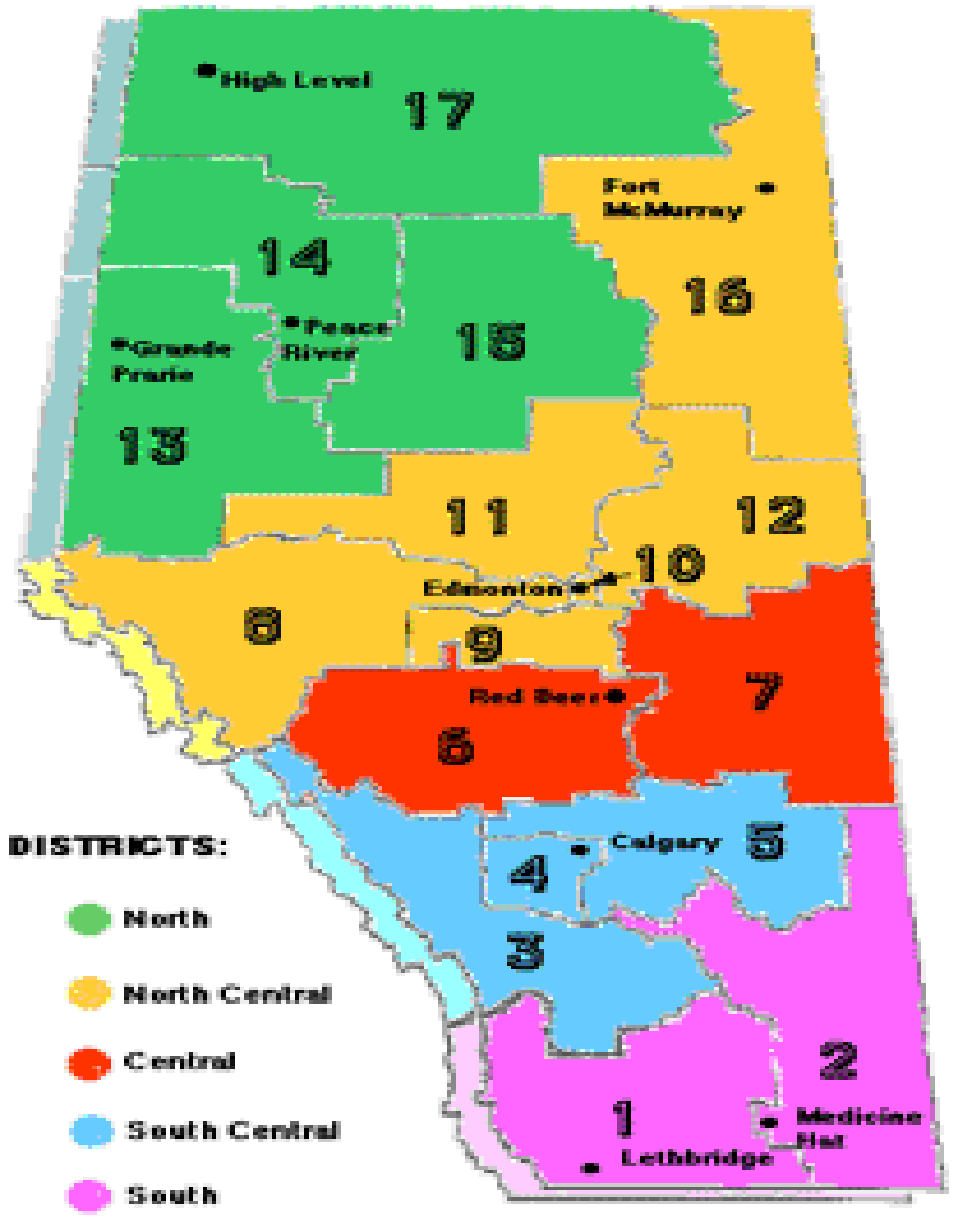
- equalizing "important" variables
  - years in school, geographic area, product use, etc.

Steps:

- Population is divided into mutually exclusive and exhaustive strata based on an appropriate population characteristic. (e.g. age, gender etc.)
- Simple random samples are then drawn from each stratum.

• Cluster Sampling

- **Divide population into clusters (usually along geographic boundaries)**
- **Randomly sample clusters**
- **Measure units within sampled clusters**



3. Selecting a Sampling Design

When to use stratified sampling

- To compare groups of different types of samples.

When to use cluster sampling

- Costly to move between each data collection location.
- There is a list of clusters but not of individual population members

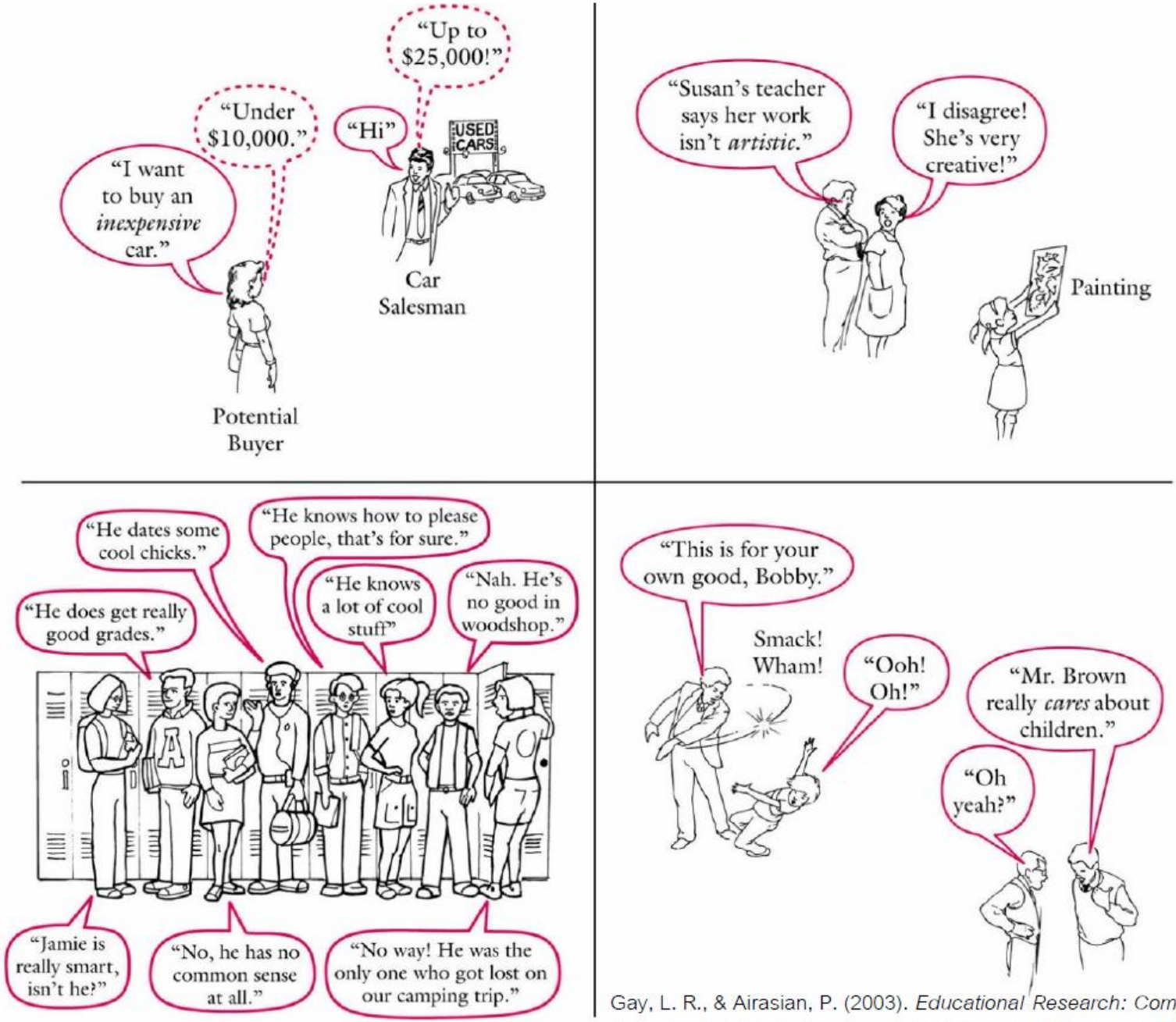
Questionnaire Design

adapted from Taylor-Powell (1998) and Burgess (2001)

1. define your research aims
2. identify the population and sample
3. decide how to collect replies
4. design your questionnaire
5. run a pilot survey
6. carry out the main survey
7. analyze the data

Questions …

and how they are perceived



Gay, L. R., & Airasian, P. (2003). *Educational Research: Competencies for Analysis and Applications*. New jersey: Merrill Prentice Hall.

Decide how to collect replies

- Interview
  - need to train all the interviewers.

- Self-administered questionnaires
  - questions have to be clear.

Questionnaire design

Collect information that answers your question.

Issue:	**Question focus:**
What reasons might candidates give for undertaking an MBA?	Is the candidate looking for: career change career advancement higher remuneration etc.
Could past experience affect the reasons?	How many years work experience does candidate have?
Could gender differences affect the reasons?	Is the candidate male or female?
Could educational background and attainment affect the reasons?	What is highest educational qualification obtained? What subject area(s) is this qualification in?
Etc.	Etc.

Layout and sequence

- Concise
- Has some introductory statement
- Logical
- Be consistent on your wordings
- Give clear instruction

e.g. *tick the relevant box*. Consider using bold for the questions themselves or for headings. Symbol fonts can be useful for characters such as ❑✓

Open vs. closed questions

Do you think football hooliganism is caused by: *(tick if appropriate)*

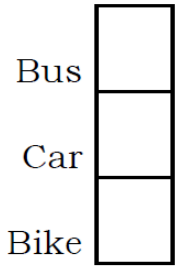
Lack of discipline at home	☐
Players' behaviour on pitch	☐
Family breakdown	☐
Youth unemployment	☐
Poor schooling	☐
Violence on T.V.	☐
Other (please specify)	☐

e.g. **What was the main problem you encountered with your wheelchair?**

☐

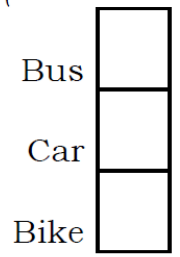
Single vs. Multiple responses

Which of the following means do you use to travel to college?



What is your <u>most usual</u> means of travelling to college?

(Tick one box only)



Complicated question

Never be tempted to use the following structure:

Select up to three of the options below and enter in the boxes opposite

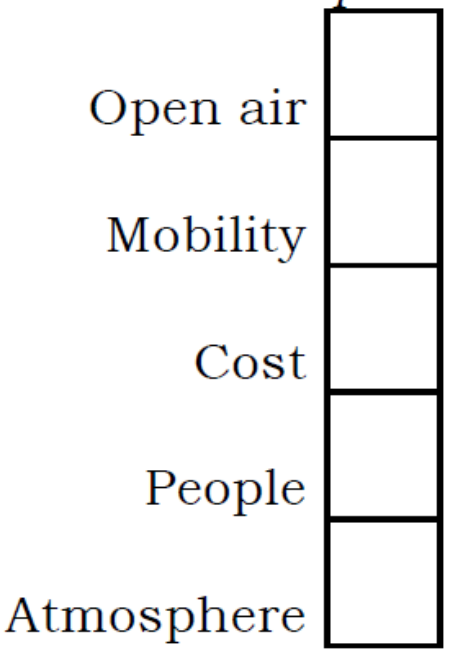
Option A	Option B	Option C
Option D	Option E	Option F

Whilst this may be logically sound you will have problems when summarising these results. A very tedious and finally unsatisfactory search is required to determine how any given option performs.

Ranked responses

Place in order of importance to you the following features of a camping holiday

(Indicate by numbering from 1-4 in order where 1 is the most important)



Rated responses

(Circle the number under the initials that applies.
VI=Very important; I=Important; N=Neutral; U=Unimportant;
VU=Very Unimportant).

Indicate your view of the following aspects of a camping holiday

	VI	I	N	U	VU
Community life	1	2	3	4	5
Low cost	1	2	3	4	5
Outdoor life	1	2	3	4	5
Ability to move around	1	2	3	4	5

Avoid double questions

Do you think the British should eat less and exercise more?

instead ask:

(Please circle relevant number)

	Yes	No
Do you think the British should eat less	1	2
Do you think the British should exercise more	1	2

Notice the consistent use of circled number responses for Yes/No questions.

Avoid questions that involve negatives

Don't confuse the respondent by language like this:

(Please circle relevant number)

	Yes	No
Are you against a ban on smoking	1	2

Ask for precise answers

Give your age on 1st September 2001 ☐☐ years

is preferable to:

Are you...

Under 18 ☐

18-65 ☐

Over 65 ☐

Avoid leading questions

Leading questions such as “Do you agree with the majority of people that the health service is failing?” should be avoided for obvious reasons that any right-minded individual can see. Don’t you agree?

Reference

Burgess (2001), “A general Introduction to the Design of Questionnaires for Survey Research”, *Information Systems Servi*ces, University of Leeds