

I. Cross-Sectional Design

A. No time dimension.

1. Can only measure differences between groups, not change.
2. Measuring change means we must look at data over time to see how individuals and/or groups actually change.
3. Since we can't do that, we do our best to infer something about change by grouping our cross-sectional data.

B. Reliance on existing differences

1. We must rely on existing differences between groups.
 - a) Cross-sectional designs are passive compared to experiments.
 - b) The “treatment” happens before we even collect the data.
2. This implies we must use control variables to control for differences between the groups caused by other variables.
3. In economics, the differences we study are measured by the dependent variable, while the controls are the independent variables.

C. Groups constructed based on existing differences.

II. Repeated Cross-Sectional Studies

A. Essentially, deVaus proposes repeating the data collection at different points in time using a different sample each time.

B. This is not quite the same as longitudinal analysis which surveys the same panel of participants over time to see how individuals change.

C. Example: census

III. Methodological Issues in Cross-Sectional Design

A. Internal validity

1. Difficult to establish causality without a time dimension.
2. Problems interpreting the results.

B. Controlling Confounding Variables

1. We use statistical controls at the data analysis stage (after the data have been collected).
2. Similar to matching groups approach.
3. However, we face the constant problem of omitting an important independent variable.
 - a) Remember, this can cause biased estimates.
 - b) Ideally you would like to include every variable you (and everyone else) can think of.
 - c) Your problem is time, cost and the inevitable tradeoff between the participant's time and their willingness to participate.
4. Independent variables that are not statistically significant can have two meanings:
 - a) They may simply not be related to the dependent variable. Eliminating possible causes of differences is almost as important as finding the causes.
 - b) Or they may be correlated with other independent variables (multicollinearity). Always look at a correlation and/or covariance matrix.
5. *A priori* versus *ad hoc* reasoning
 - a) We want to use *a priori* reasoning, developing a model and testing hypotheses implied by the model.
 - b) Beware of *ad hoc* reasoning which develops hypotheses based on the data. You can always come up with a story that explains statistical results. That's not good research.

C. Establishing Causal Direction

1. Correlation does not mean causation.
2. Theory establishes causation.
3. We test our theory with statistical techniques using our data.
4. There remains the problem of inferring causal direction. This is just about impossible with cross-sectional data.

D. Interpreting Results

1. Must show that correlation is the result of choices made by the actors in the study.
2. Your theory should imply choices.
3. Your data must measure the choices made.
4. deVaus and others argue that this loses the “human” dimension of the problem.
 - a) Some people (none economists) try to use a holistic approach.
 - b) Economists don’t do this, as it usually results in a case study.

IV. External Validity

A. Representative Sample?

1. Cross-sectional designs should be representative because you can look at your sample as you’re collecting the data.
2. No problems with sample attrition, mortality, and other problems that arise in time series studies.

B. Weighted Samples

1. Problem: n may be small in some cells.
2. Solution: oversample those cells, then correct your final results to match the population distribution.

V. Practical Issues

A. Method of collecting data

1. Set up a “variable by case” matrix (Fig. 11.2).
2. Sample must have variation in each variable.
3. Can collect the data using questionnaires, face-to-face interviews, observations, telephone interviews, and so on.
 - a) deVaus says you can mix the methods for a single study.
 - b) Must use the same mix of methods for each case.
 - c) Problem: how do you recognize the “case” before you’ve asked any questions?
 - d) Advice: stick to a single method

	Face-to-face	Telephone	Mail
<i>Response rates</i>			
General samples	Good	Good	Good
Specialized samples	Good	Good	Good
Representative samples			
Avoidance of refusal bias	Good	Good	Poor
Control over who completes the questionnaire	Good	Good	Satisfactor y
Gaining access to the selected person	Satisfactory	Good	Good
Locating the selected person	Satisfactory	Good	Good
<i>Effects on questionnaire design</i>			
Ability to handle:			
Long questionnaires	Good	Satisfactory	Satisfactor y
Complex questions	Good	Poor	Satisfactor y
Boring questions	Good	Satisfactory	Poor
Item non-response	Good	Good	Satisfactor y
Filter questions	Good	Good	Satisfactor y
Question sequence control	Good	Good	Poor
Open ended questions	Good	Good	Poor

<i>Quality of answers</i>			
Minimize socially desirable responses	Poor	Satisfactory	Good
Ability to avoid distortion due to:			
Interviewer characteristics	Poor	Satisfactory	Good
Interviewer opinions	Satisfactory	Satisfactory	Good
Influence of other people	Satisfactory	Good	Poor
Allows opportunities to consult	Satisfactory	Poor	Good
Avoids subversion	Poor	Satisfactory	Good
<i>Implementing the survey</i>			
Ease of finding suitable staff	Poor	Good	Good
Speed	Poor	Good	Satisfactory
Cost	Poor	Satisfactory	Good

Source: adapted from Dillman, 1978

B.

Table 11.1 Advantages and disadvantages of three methods of questionnaire administration

Sample Size

- 1. Depends on funds, time, access to participants, planned methods of analysis and the degree of precision and accuracy required.**
- 2. The larger the sample the better.**
- 3. You can easily reduce sample size by designing a long, complex questionnaire that requires 30 minutes of participant time to complete. (I recently hung up on a telephone interviewer because the survey took too much time.)**

C. Precision and Accuracy of Estimates

- 1. Larger sample means more precision.**
- 2. Table 11.2 shows this.**

Table 11.2 Sample sizes required for various sampling errors at 95 per cent confidence level (simple random sampling)

ing 1	Sam- ple size ²	Sam- pling error ¹ (%,)	Sam- ple size ²
	9,200	5.5	330
	4,500	6.0	277
	2,400	6.5	237
	1,600	7.0	204

¹This is in fact two standard errors.

²This assumes a 50/50 split on the variable. These sample sizes would be smaller for more homogeneous samples.

**3.
Example: most national surveys reported in the press have a sample size of 625 individuals, achieving the often-reported error of $\pm 4.0\%$**

D. Optimal Sample Size

1. The optimal sample size is the size at which marginal cost equals marginal benefit.
2. If you have a large budget, marginal cost may be low.
3. If accuracy requires a larger n in some cells, the marginal benefit may be high.

E. Variation in Key Variables

1. We need variation in the dependent variable.
2. We also need variation in the independent variables. Otherwise, they won't be able to explain much.

F. Statistical Controls

1. Since the controls will be applied after the data is collected, you must think about what controls you will need before you design the survey instrument.
2. Once you've thought about them, be sure to include them in your survey instrument.

G. Length

1. Questionnaires can't be too long or your response rate will suffer.
2. However, if your topic is of interest to people, you can get away with a somewhat longer survey instrument.

H. Types of Data

1. Must choose between open ended and forced choice responses.
2. Must also choose between nominal and ordinal choices.
 - a) Economists generally prefer nominal data because ordinal data rapidly uses up degrees of freedom.
 - b) However, some data (e.g., gender) must be ordinal.

VI. Cross-Sectional Analysis: Descriptive Phase

A. Counting

1. We count responses to different questions because we want to look at the distribution of responses.
2. Where population data is available, response frequency should be compared to population frequency as one measure of how representative your sample is.

B. Collapsing interval variables

1. Be careful doing this. The way variables are collapsed can change your interpretation of the results.

2. Table 12.1 (p. 197) is an excellent illustration of this.
3. One solution: let the distribution of the data determine your intervals rather than assigning them yourself.

C. Form of data

1. In general, the statistical software will handle problems in this area.
2. However, you may want to rescale a variable to make its coefficient easier to interpret.

3. deVaus suggests normalizing variables, viz., $\frac{\bar{x} - \mu}{\sigma}$.

VII. Cross-sectional Analysis: Explanatory Analysis

A. Statistical Controls

1. The *ceteris paribus* assumption is embodied in our independent variables.
 - a) A regression coefficient tells us how much our dependent variable will respond to a change in one independent variable holding all other independent variables constant.
 - b) One common procedure is to set the other independent variables to their mean values, then look at the response of the independent variable. For example, elasticities are often calculated at the means of the variables.
2. "... it is not possible to control for every possible variable, so the possibility always remains that any ... differences could be due to these uncontrolled variables." (deVaus, p. 203)
 - a) This corresponds exactly to the "omitted variable" problem in statistics.
 - b) While it is important to discuss this problem, if you worry about it too much you'll never get any research done.

VIII. Some Statistical Concerns

A. Heteroscedasticity

1. This is often a problem in cross-sectional analysis
2. Three common tests; use groupwise heteroscedastic model (see William H. Greene, Econometric Analysis, Macmillan, 1993, sections 14.3.1, 14.3.4, and 16.3.1):

$$V = \begin{bmatrix} \sigma_1^2 I & 0 & 0 & 0 \\ 0 & \sigma_2^2 I & 0 & 0 \\ 0 & 0 & \sigma_3^2 I & 0 \\ 0 & 0 & 0 & \sigma_4^2 I \end{bmatrix}$$

In other words, allow σ^2 to vary across i , constructing groups. In the example shown above there are four groups. Then perform one of the following tests.

a) Lagrange multiplier test

$$LM = \frac{T}{2} \sum_i \left[\frac{s_i^2}{s^2} - 1 \right]^2$$

where T is the number of time periods (equal to 1 for pure cross-sectional data). The test is a chi square test with n degrees of freedom.

b) White's test

Regress the squared OLS residual on a constant and all combinations of the independent variables, viz., X_1 , X_2 , X_1^2 , X_2^2 , and X_1X_2 . Calculate the chi-squared statistic as $(nT)R^2$ and perform the usual chi squared test. Note that the null hypothesis being tested is that the data is homoscedastic. Rejecting the null hypothesis means you have a heteroscedasticity problem.

While White's test is very robust and doesn't assume normality (which the LM test does), it chews up degrees of freedom rather rapidly.

c) Approximate likelihood ratio test

The approximate likelihood ratio statistic is

$$-2 \ln \lambda = (nT) \ln \hat{\sigma}^2 - \sum_i T \ln \hat{\sigma}_i^2$$

where $\hat{\sigma}^2 = \frac{e'e}{nT}$ and $\hat{\sigma}_i^2 = \frac{e'_i e_i}{T}$

This chi-squared statistic has $n-1$ degrees of freedom. If only least squares are available, s^2 and s_1^2 may be used. However, you will lose some power of the test, particularly if your sample is small.

3.

B. Sample size is always a concern, particularly when working with primary data. Remember, you lose one degree of freedom for each independent variable. If you have k independent variables, you should have at least $50+k$ observations.

IX. Issues in Questionnaire Design

A. The standard (and only) reference is William Foddy, Constructing Questions for Interviews and Questionnaires: Theory and Practice in Social Research, Cambridge University Press, 1993.

B. Causes of error in gathering data using surveys (p. 2)

- 1. Respondents' failure to understand the questions as intended.**
- 2. A lack of effort, or interest, on the part of respondents.**
- 3. Respondents' unwillingness to admit to certain attitudes or behaviors.**
- 4. Failure of respondents' memory or comprehension processes in the stressed conditions of the interview.**
- 5. Interviewer failures of various kinds (e.g. the tendency to change wording, failures in presentation procedures and the adoption of faulty recording procedures).**

C. Examples of causes of errors

- 1. Factual questions sometimes elicit invalid answers.**
- 2. The relationship between what respondents say they do and what they actually do is not always very strong.**
- 3. Respondents' attitudes, beliefs, opinions, habits, interests often seem to be extraordinarily unstable.**
- 4. Small changes in wording sometimes produce major changes in the distribution of responses.**
- 5. Respondents commonly misinterpret questions.**
- 6. Answers to earlier questions can affect respondents' answers to later questions.**
- 7. Changing the order in which response options are presented sometimes affects respondents' answers.**
- 8. Respondents' answers are sometimes affected by the question format *per se*.**

9. Respondents often answer questions even when it appears that they know very little about the topic.

10. The cultural context in which a question is presented often has an impact on the way respondents interpret and answer questions.

D. The key issue: the comparability of answers (p. 17)

1. The researcher must be clear about the nature of the information required and encode a request for this information.

2. The respondent must decode this request in the way the researcher intends it to be decoded.

3. The respondent must encode an answer that contains the information the researcher has requested.

4. The researcher must decode the answer as the respondent intended it to be decoded.

E. Symbolic interactionist theory (Blumer, summarized in Foddy pp. 19 ff.)

1. Human beings interpret and define each other's actions.

2. Human beings can be the objects of their own attention. In other words they can act toward themselves as they act toward others.

3. Conscious social behavior is intentional behavior.

4. Interpreting, planning and acting are ongoing processes which begin anew at every stage of a social interaction. Both parties in a dyadic interaction engage in these processes.

5. Human intelligence is, in part, reflexive in character.

6. These processes occur in all social situations (although they will be most obvious in newly formed situations as the interactants struggle to align their behaviors with one another).

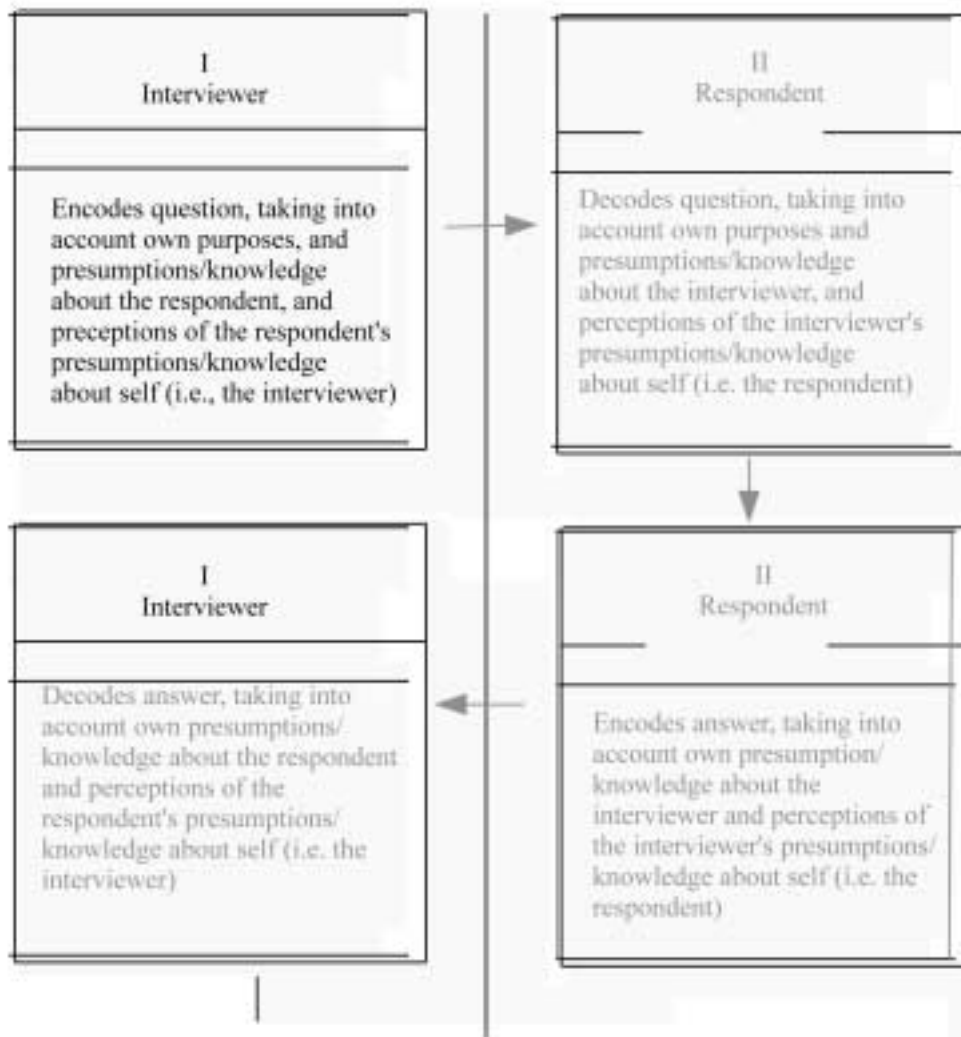


Figure 2.2 in Foddy (p. 22)

X. Conclusion

A. Cross-sectional data analysis avoids some (but not all) statistical problems.

B. Make sure you have a large enough sample size.

C. Design your questionnaire correctly.

