M4903-450
Quantitative Methods in Economics

Part II: Quantitative Research Designs and Econometrics

Prof. Dr. Manfred Zeller
Dr. Alwin Keil

Fachgebiet Entwicklungstheorie und -politik für den ländlichen Raum (490a),
University of Hohenheim, Schloß, Osthof-Süd, 70599 Stuttgart
www.uni-hohenheim.de/i490a
E-mail: manfred.zeller@uni-hohenheim.de,
alwin.keil@uni-hohenheim.de

Overview of the module

- Phases of socio-economic research
- Quantitative research designs in the social sciences
- Constructing sampling frames, and different sampling procedures
- Questionnaire design and data entry
- Data cleaning: Handling outliers and missing values
- Parametric and non-parametric tests
- Linear regression: The Gauß-Markov Theorem
- Models correcting for sample selection bias

Major references (1)

- **Bernard, Russell H. (2000)** Social research methods. Qualitative and quantitative approaches. Sage Publications, London.
- **Black, Thomas R. (1999)** Doing quantitative research in the social sciences. An Integrated approach to research design, measurement and statistics. Sage Publications, London.
- **Burns, Robert B. (2000)** Introduction to research methods. Forth Edition. Sage Publications, London.
- **Deaton, Angus (2000)** The analysis of household surveys : a microeconometric approach to development policy. Johns Hopkins University Press, Baltimore and London.

Major references (2)

- **Field, Andy (2005)** Discovering statistics using SPSS. Second Edition. Sage Publications, London.
- **Gujarati, Damodar N. (2003)** Basic econometrics. Forth Edition. McGraw-Hill, New York.
- **Hill, R. Carter, Griffiths, William E., and Judge, George G. (2001)** Undergraduate econometrics. Second Edition. John Wiley & Sons, New York.
- **Kennedy, Peter (2003)** A guide to econometrics. Blackwell, Massachusetts.
- **Maxim, Paul S. (1999)** Quantitative research methods in the social sciences. Oxford University Press, New York.

Major references (3)

- **Minten, Bart, and Zeller, Manfred (2000)** Beyond market liberalization: Income generation, poverty alleviation, and environmental sustainability in Madagascar. Ashgate Publishing Company, Aldershot.
- **Sadoulet, Elisabeth, and de Janvry, Alain (1995)** Quantitative Development Policy Analysis. John Hopkins University Press, Baltimore and London.
- **Statistical Package for the Social Sciences (SPSS).** Various manuals, 1999/2000
- **Tatian, Peter (1992)** Designing a data entry and verification system. Microcomputer in policy research series NO. 1. IFPRI, Washington D.C.

Qualitative and quantitative research as complements (Black, 1999)

Use of design depends on research problem, and the two research designs (RDs) complement each other

Single or a few selected groups (case study): *Why?* *How?*	⇄	Representative sample(s): *What is?* *How widespread?* *Why (causal research)?*

Types of research designs (RDs) (1)

Qualitative RDs: use a case study approach to determine observation units and use field research instruments such as: Participant observation, unstructured in-depth interviews, participatory rural appraisal

→ few observation units researched in detail

Types of research designs (2)

Quantitative RDs: use representative samples and mainly use structured questionnaires with standardized definition of variables that are to be measured

→many observation units researched about narrowly defined issues for which results are sought to be representative

Use of design depends on research problem, and the two RDs complement each other!

Assumptions of qualitative versus quantitative Research Designs (RDs) (Source: Burns, 1999)

Qualitative

- Variables complex and interwoven; difficult to measure
- Events viewed from informant's perspective

Quantitative

- Variables can be identified and measured
- Events viewed from outsider's perspective

Purposes of qualitative versus quantitative RDs (1)

Qualitative

- Understanding the perspectives of others
- Interpretation
- Contextualisation

Quantitative

- Causal explanation
- Prediction
- Generalisation

Methods of qualitative versus quantitative RDs (2)

Qualitative

- Data collection using participant observation and unstructured interviews
- Concludes with hypothesis and grounded theory

Quantitative

- Measuring and testing
- Commences with hypothesis and theory

Methods of qualitative versus quantitative RDs (3)

Qualitative

- Inductive
- Data analysis by themes from informants' descriptions
- Data reported in language of informant

Quantitative

- Deductive
- Statistical analysis
- Statistical reporting

Role of researcher in qualitative versus quantitative RDs (Source: Burns, 1999)

Qualitative	Quantitative
• Researcher as instrument	• Researcher applies formal instruments
• Personal involvement	• Detachment
• Empathic understanding	• 'Objective'

Strengths of qualitative RDs (Source: Burns, 1999) (1)

The value of qualitative studies lies in their ability to research issues that:

- explore folk wisdom, cultures and traditions
- investigate real or hidden agendas of organisations as opposed to stated agendas
- cannot be done experimentally (for ethical or practical reasons)
- unravel informal and unstructured links and processes in organisations
- delve in depth into processes

Strengths of qualitative RDs (2)

→ **Strength of qualitative studies then lies in research that is descriptive or exploratory and that stresses the importance of context and the subjects' frame of reference.**

Quantitative research designs: Stages of the research process

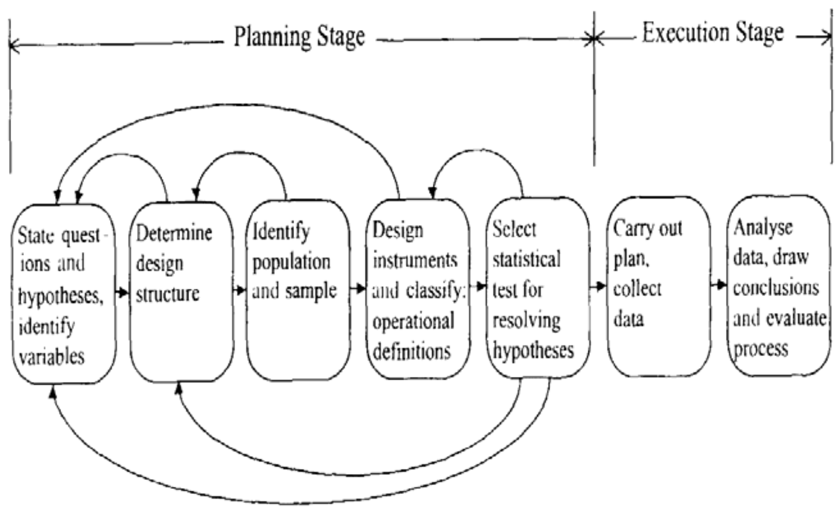


FIGURE 2.1 Stages of designing and carrying out a study, including iterations for modifications and improvements during planning

Source: Black, 1999: 27

Developing and testing theories in social sciences

- **Theory** = formal statement of how variables relate to each other (distinguishing a dependent variable Y and independent variables Xs)
- Common method of **developing a theory**: *Induction* = Observations are made, data are collected, general patterns are recognized, and relationships between variables are proposed that form a theory
- Common method of **testing a theory**: Qualitative and quantitative research rejecting or accepting *hypotheses* postulated by theory

(See Figure 1.2 in Black, 1999)

Stage 1: Stating research questions and hypotheses (1)

Research question

→ Hypothesis(es)

→ Null-hypothesis(es)

Stage 1: Stating research questions and hypotheses (2)

- When it comes to testing hypotheses, all that statistics can tell us is whether the outcomes we ultimately see could have happened due to some causal relationship or simply by chance.
- A null-hypothesis hereby simply states that ‚no significant difference‘ is expected between what we observe as difference in the data and what would happen by chance alone. If the difference is greater than some minimum, it is considered statistically significant, and whatever has happened (probably) did not occur by chance alone.

Stage 1: Stating research questions and hypotheses (Example 1)

- **Research question** (R1): Are there any relationships between education level of a mother, location of her home and the vaccination status of her children?
- A possible **hypothesis** (H1): It is expected that there will be a negative relationship between distance to clinic and child vaccination, and a positive relationship between education and child vaccination.
- **Null-hypothesis** (H0) for statistical test: There will be no significant correlation between distance to clinic and vaccination, or between mother‘s education and vaccination status for a random selection of mothers.

Stage 1: Stating research questions and hypotheses (Example 2)

- **R1**. Which types of in-kind loan packages for agricultural inputs had a greater success in increasing hybrid maize production in the last rainy season?
- A possible **H1**:. One of three loan packages (A, B, or C) received by farmers will result in a higher share of hybrid maize in total area cultivated by farmers during the last rainy season.
- **H0** for statistical test: There will be no significant difference in the share of hybrid maize in total cropped area across a random sample of farmers who received a loan from package A, B or C.

Types of quantitative research designs (1)

All quantitative research designs use **samples** of the population. If taken randomly, these samples can be representative of the population

Types of quantitative research designs (2)

Correlational (or descriptive) research	Causal research
→ asks, whether there exists a relationship between two variables *Example:* *Is the number of extension messages received related to the adoption of improved wheat seeds? In other words: do the variables correlate?*	→ asks whether a certain variable, often denoted X, influences another variable (Y). *Example:* *Is there a difference in adoption of improved wheat seeds between farmers who received extension messages and those who did not?* *Does extension (X) influence adoption (Y)?*

Types of designs for causal research (Source: Black, 1999)

A. **Pre-experimental** (no potential for causality, purely descriptive, do NOT use this design if we wish to address causal research questions. However, it is presented here to learn from the mistakes)

B. **Experimental** (the ideal design for testing causality of hypothesized relationship X → Y, but mostly not practicable in social science research, especially in developing countries)

C. **Quasi-experimental** (sometimes used, better than D in establishing causality)

D. **Ex-post facto** (after the fact, after the treatment/policy has been applied) → most often used for causal socio-economic research

Types of designs for causal research: Notation used

Notation	Meaning
X_a, X_b, X_c, …	Potential **independent variables** (‚treatments'), and
X_0	a **‚control'** treatment where nothing happens.
O	**Observation** made on subjects before/after treatment,
O_1, O_2, …	with more than one observation made on the group of subjects
O_{a1}, O_{a2}, .. , O_{b1}, ..	or multiple observations on different treatment groups of subjects.
Y	Potential **dependent variable**, which could be operationally defined as O or some combination of O_1, O_2, …, such as gain scores ($O_2 – O_1$),
Y_a, Y_b, Y_c, …, Y_0	with the subscript indicating the value of Y corresponding to the category of the independent variable X_a, X_b, X_c, … (see above).
RS	**Random selection** of a single sample from a larger population, followed by
RA_a, RA_b, ..., RA_0	**random assignment** of subjects to experience treatments as the independent variables X_a, X_b, X_c, … (manipulated), or
RS_a, RS_b, …, RS_0	**random selection of several samples** where 'life experiences' are the independent variables X_a, X_b, X_c, … (observed).
(RS)	In a design, this indicated that random selection is optional and unlikely to make a difference.

Source: Black, 1999: 65

Pre-experimental research design (1)

A1: One group post-test /observation

Design structure (time sequence of events):

(RS) → X → O_1

One group tested or observed for a dependent variable Y for one instance of the independent variable X, which is itself either observed or manipulated. Y is measured by the variable O which is the operational definition of the dependent variable Y (O is the empirical measure used as instrument to measure Y).

Pre-experimental research design (2)

Example: Wheat farmers that have received extension service (X) are surveyed, and we ask whether they adopted a certain improved wheat variety or not (the O).

Problem: A single group (say wheat farmers having attended an extension meeting talking about improved seeds), nothing equivalent with which to compare, even if the group was randomly selected → results can only be descriptive, no causality between X and Y can be established (e.g.: X = Attendance of extension meeting, Y = adoption of modern seed varieties for wheat). For example, those farmers that have adopted may have done so because the price of wheat increased.

Pre-experimental research design (3)

A2: One group pre- and post-test /observations

Design structure:

(RS) → O_1 → X → O_2

One group is observed before some single event (or treatment) (the **pre-test**) and then thereafter (the **post-test**). Measurements of a certain outcome variable O are taken before and after the event. The event or treatment is the variable X of interest. The difference between the pre- and post-measure of O is indicative of the dependent variable Y ($Y = O_2 - O_1$).

Pre-experimental research design (4)

Example: We would measure the adoption of improved seed in a random sample of wheat farmers before any extension was offered. After a couple of extension meetings (and growing seasons), we would go back to the same group of farmers, and ask again if they had adopted (before: 10 % of farmers, afterwards: 25 % of farmers). We would accrue the difference in adoption rates Y (15 %) as the effect of the extension.

Pre-experimental research design (5)

Problem: There is a range of *extraneous variables* that could have also caused this observed difference. For example: improved seeds became much cheaper, fertilizer became cheaper, or effective radio messages about the wheat varieties helped the adoption, etc.

→ No causality can be established if these so-called extraneous variables are not controlled!

Pre-experimental research design (6)

A3: Two-group post-test /observations

Design structure:

$X_a \rightarrow O_{a,}$ and $X_0 \rightarrow O_0$

Two groups are chosen **purposely (not randomly)**. One group (denoted as a) was subjected to a policy/treatment (or experienced an event), the other was not (control = denoted as 0). Measurements of a certain outcome variable O are taken in the two groups after the event/policy/treatment, where O is an indicator of Y. The event, policy or treatment is the variable X of interest. The difference for O between the two groups ($O_a – O_0$) is interpreted as the effect of X on Y.

Pre-experimental research design (7)

Example: We would measure the adoption of improved seed in a group of purposefully selected wheat farmers who received the extension message. The so-called control group consists of farmers that did not receive the extension (whereby other characteristics are similar). We may find that 15 % of those with extension have adopted, and only 8 % of those without extension. We would (falsely) accrue the difference in adoption rates (7 %) as the effect of the extension (X) on adoption (Y).

Pre-experimental research design (8)

Problem: There is a range of *extraneous variables* that could also have caused this difference. For example, the group receiving the message may have had better education → questioning whether the groups have the same status.

→ A PURPOSEFULLY SELECTED SAMPLE CAN NEVER CONTROL FOR EXTRANEOUS VARIABLES!

Experimental research design (1)

True experiments always have two groups: one group is subjected to the treatment/policy, the other is not (The latter is called the control group). Preferably, the observation units are randomly assigned to the two groups (denoted as RA) and in both groups, random samples (denoted as RS) are taken.

Experimental research design (2)

A1 (one group)
→ no equivalent true experiment
A2 (one group pre- and post-test)
→ extended to two groups pre- and post-test (so-called B2)
A3 (two purposefully selected groups with post-test)
→ two groups, but observation units are randomly selected, with post-test (so-called B3)

Experimental research design (3)

B2: Two group pre- and post-test /observations

Design structure:

$$RS \rightarrow \begin{cases} RA_a \rightarrow O_{a1} \rightarrow X_a \rightarrow O_{a2} \\ RA_0 \rightarrow O_{01} \rightarrow X_0 \rightarrow O_{02} \end{cases}$$

Experimental research design (4)

A random sample (RS) of the population is taken, and randomly assigned to two groups. One group is subjected to the event/treatment or policy, the other is not. The two groups are measured before and after the treatment (with respect to the outcome variable O). The dependent variable is defined by the *gain scores*, i.e. the difference in scores between the two measurements for a group:

$Y_a = O_{a2} - O_{a1}$ and $Y_0 = O_{02} - O_{01}$.

The effect of X is: $Y_a - Y_0$.

Experimental research design (5)

Example: We would measure the adoption of improved seed in a random sample of wheat farmers who were randomly assigned to two groups (e.g. Group *a* has 8 %, and Group *0* has 9 %). The experiment group *a* would receive the extension, while the other would not. In the next growing season, we would go back to the same two groups of farmers and ask again about adoption, and find e.g. that Group *a* has 25 % of adopting farmers and Group *0* only 13 %...

Experimental research design (6)

…We could now test whether the difference in change in adoption (25 – 8 = 17% versus 13 – 9 = 4%) between the two groups is statistically significant (not just a difference that could have happened by chance).

→ This is the almost **ideal research design**. This design is widely used in natural science (agronomy, medicine etc.).

Experimental research design (7)

Problems: This design is, however, seldom used in social science research because the potential for **control over experimentation** often does not exist. Sometimes, social science research may also not have **control over sampling** the two groups appropriately because of ethical reasons, time or budget reasons. In our example, all farmers assigned to the treatment group MUST go to the extension meeting, even if they do not wish to. The MUST may cause lots of problems (ethical, social tensions, etc.)…

Experimental research design (8)

…The potential for control over experimentation exists for certain projects (i.e., the ‚treatment'), such as extension/credit/education, but most projects do not take random baseline surveys; and they do not randomly assign groups that receive the project treatment and those that do not. However, even if the treatment and the control group were randomly assigned, one would still need to rule out that one group learns about the other (for example by randomly selecting different villages that are far apart)...

Experimental research design (9)

…However, if we were to choose different villages, the different agronomic and socio-economic characteristics of these villages may result in non-comparable samples of farmers (i.e., thus actually employing a quasi-experimental design, see below). We also would need to assure that all wheat farmers who were chosen to receive the treatment actually receive it (e.g., by paying them for attendance to the meeting). However, paying may limit the value of this extension as an example for real-life situations…

Experimental research design (10)

…The potential for control over experimentation does not exist for so-called life-events, for example being born in a wealthy family, being male, living in a city, being a migrant. These life events are important independent variables (*X*) in social science research. The potential for control over experimentation is also limited for nation-wide policies, such as price changes or changes in trade policies. Here, everybody is potentially affected, i.e., receives the treatment, there is no control group.

→ **It is very difficult (almost impossible) to conduct true experiments in social science!**

Experimental research design (11)

B3: Two-group post-test /observation

Design structure:

$$RS \rightarrow \begin{cases} RA_a \rightarrow X_a \rightarrow O_a \\ RA_0 \rightarrow X_0 \rightarrow O_0 \end{cases}$$

A random sample (RS) of the population is taken and randomly assigned to two groups. One group is subjected to the event/treatment or policy, and the other is not. Measurements of the outcome variable O in the two groups are only taken after the treatment. The dependent variable is defined by the difference in scores between the two groups:

$Y_a = O_a$ and $Y_0 = O_0$. The effect of X is measured as $Y_a - Y_0$.

Experimental research design (12)

Example: We would select a random sample of wheat farmers who then are randomly assigned by us to two groups. The experiment group *a* would receive extension on improved wheat seed, while the other would not. In the growing season following the extension service, we would determine adoption in both groups of farmers. We may find e.g. that Group *a* has 25 % of adopting farmers and Group *0* only has 13 %. We could now test whether the difference in adoption is statistically significant, i.e., not just a difference that could have happened by chance.

Experimental research design (13)

Compared to A3, we ensure comparability between the two groups through *randomization* (taking random samples of wheat farmers and randomly assigning them to two groups). Thus, all extraneous variables (such as age, gender, education, etc.) are controlled for, because members of the two groups should not differ with respect to the extraneous variables.

→Because of this, there is much stronger support for causality: X → Y.

Experimental research design (14)

Problems: Not having a pre-test (like B2) has advantages and disadvantages. The advantage is that the pre-test will not become a variable that potentially influences the subjects before treatment (e.g., the wheat farmers before they receive extension). The disadvantage is that one does not have a measure beforehand with which to compare. This is NOT a problem in situations where we can be sure that the random assignment resulted in two comparable groups…

Experimental research design (15)

…However, in social science research, criticism can be often raised that the two groups may in fact differ in certain characteristics. Nonetheless, B3 is a useful experiment in social science research that provides **reasonable support for causality if randomization is carefully done**.

Quasi-experimental research design (1)

Quasi-experimental designs employ strategies that have a similar structure to experimental designs, but **lack much of the control of a true experimental design**. The researcher has control over data collection (when?, from whom?), but lacks complete control of the treatment (when and to whom the treatment will be exposed).

The when?-question may be determined by project or government officials, not by the researcher. Related to the whom?-question, the definition of the treatment group in quasi-experiments **does not rely on a random assignment**, but either:

Quasi-experimental research design (2)

- on **self-selection** (i.e., the voluntary choice of people to belong to a group, e.g., farmers decide whether they attend an extension meeting about wheat seed or not)

or

- on **administrative decisions** (e.g., the government/project offers the extension service only in villages close to the road).

Quasi-experimental research design (3)

Because of these two options, the **treatment and the control group may not be comparable** because they may differ in extraneous variables that could influence Y.

In the quasi-experiment, the researcher may still control the nature of the experiment (provided that he or she can cooperate with project/government etc.), but the subjects tend not to be randomly selected or randomly assigned, nor do they necessarily constitute equivalent groups for all characteristics.

Quasi-experimental research design (4)

→ **Quasi-experiment: Control over data collection, partial control over the experiment, but no control over randomly assigning people to groups!**

Quasi-experimental research design (5)

C2: Quasi-experimental design with pre-test

Design structure:

$RS_a \rightarrow O_{a1} \rightarrow X_a \rightarrow O_{a2}$

$RS_0 \rightarrow O_{01} \rightarrow X_0 \rightarrow O_{02}$

Quasi-experimental research design (6)

Most often (under field conditions of social science research) it is impossible to randomly assign the treatment and the control group. The treatment and the control group is not determined by random assignment but by self-selection or administrative decisions. However, **random samples of observation units belonging to these groups are selected**. The above quasi-experiment measures the outcome variable before and after the experiment in both groups…

Quasi-experimental research design (7)

…The treatment group is subject to the event/treatment or policy, and the other is not. The dependent variable is defined by the difference in gain scores between the two groups:

$Y_a = O_{a2} - O_{a1}$ and $Y_0 = O_{02} - O_{01}$.

The effect of X is $Y_a - Y_0$.

Quasi-experimental research design (8)

The first measurement is often called the *baseline*, and well-planned development projects undertake a baseline survey for purposes of being able to assess the effect of the project later. Then the project offers a service (e.g. extension, credit etc), and people *self-select* themselves into either using this service (then they belong to the treatment group) or not (the control group).

Quasi-experimental research design (9)

Example: We are interested in the adoption of improved wheat seed due to extension. Before the extension is offered, we take a random sample of wheat farmers in a (preferably randomly) selected village of our project area and assess adoption. This is the baseline. Then, the project/government offers an extension service in that village. Farmers may choose to listen to the extension message or not (they self-select themselves into either the treatment or control group)...

Quasi-experimental research design (10)

…After the extension service has been offered, we revisit the same sample of farmers and ask whether they have attended the extension meeting (if yes, they belong to our treatment group), and whether they grow improved wheat varieties. We may find that among the treatment group *a* adoption has increased by 25% and among Group *0* only by 8%. We could now test whether the difference in increase in adoption (25% versus 8%) between the two groups is statistically significant, i.e., not just a difference that could have happened by chance.

Quasi-experimental research design (11)

Problems: The above practice is likely to underestimate the true effect of extension because of learning of farmers in the control group from the farmers in the treatment group. How to control this? → Select many villages randomly (including ones with no treatment), and only in some conduct (or observe) the extension pilot project.

Quasi-experimental research design (12)

The effect would then be measured as the **sum of two effects**:

- Y_{aT} - Y_{0T} (**Initial extension effect** measured in treatment villages (T) only)
- Y_{0T} - Y_{0C} (**Cross-learning effect** measured among members of the control group in treatment villages T and control villages C)

This design was used by Pitt et al. (Journal of Political Economy, 1999) in assessing the impact of Grameen Bank on a range of income and welfare variables. The sample had approx. 2000 survey households in about 80 villages !

Quasi-experimental research design (13)

Other problems of quasi-experimental design: Because the treatment and the control groups cannot be randomly assigned, the two groups may differ in extraneous variables influencing Y because of self-selection or administrative assignment (case can be strenghtened if we can show that the two groups do not differ in Y in the pre-test). In the case of self-selection, the **selection bias** can be potentially controlled for by two-stage Heckman regression models (which are available in LIMDEP or STATA stat routines).

Strengths of quasi-experiment: External validity → generalizability to real-life situations

Ex-post facto research designs (1)

In social science, many independent variables of interest tend to be natural or life experiences, such as level of education, belonging to a certain social class, village or nation, location of residence, occupation, etc. One cannot conduct an experiment to analyze the influence of such variables on some outcome variable Y, but can only observe Xs and their relationships with Ys:

The above Xs cannot be controlled directly by the researcher for practical, ethical or other reasons. **They can only be observed by the researcher after the fact (therefore ex-post facto).**

Ex-post facto design (2)

For reasons of practicality, ex-post facto research designs are the most widely used for causal social science research in developing countries.

→ For ex-post facto designs, **random selection is extremely important** in order to:

- make inferences about relationships between Xs and Ys for the general population
- and to have some chance of making statements about causal relationships

Ex-post facto design (3)

D3: Post-test /observation only, with a control group

Design structure: $RS_a \rightarrow X_a \rightarrow O_a$

$$RS_0 \rightarrow X_0 \rightarrow O_0$$

In this type of ex-post design, a random sample of the population is taken. The sample may be *stratified* by some independent variables (but not necessarily so), such as whether farmers received extension or not, whether the villages they live in are close to the road, etc. Such stratification may lead to the definition of a treatment group and a control group (denoted as *a* or as *0* in above design structure).

Ex-post facto design (4)

In any case (simple or stratified random sampling), the survey will measure X and O. Based on the information contained in X, the sample can be divided into two groups. The dependent variable is defined by the difference in scores between the group which experienced a certain X (the life event of interest, i.e., Group *a*) and the group that did not (Group *0*):

$Y_a = O_a$ and $Y_0 = O_0$.
The effect of X is measured as $Y_a - Y_0$.

Ex-post facto design (5)

Example: We would measure the adoption of improved seed in a random sample of wheat farmers. The population of wheat farmers in a certain area would be stratified into two groups: One that received extension, and one that did not. In each of the two groups, we would select a random sample of farmers and determine the adoption rate. We may find e.g. that Group *a* has 25 % of adopting farmers and Group *0* only 13 %. We could now test whether the difference in adoption (25% versus 13%) between the two groups is statistically significant, i.e., not just a difference that could have happened by chance.

Ex-post facto design (6)

Problems: However, we CANNOT say (even in the case of statistical significance using a chi-square test) that extension caused this to happen because those farmers who received extension may differ in many other **extraneous variables** from those who did not (e.g., access to markets, education, farm size → all these variables may influence adoption). For causal research using ex-post designs, we need to control for potential extraneous variables through **regression analysis**. For example, apart from extension, our theory/conceptual framework suggests that education is also a potential determinant of adoption.

Ex-post facto design (7)

We would need to also measure education in the sample. Note that this is also a life event variable. In addition, the farm size may be a variable in our conceptual framework. A regression model can estimate the relationships between each and every X and the Y by controlling for the influence of the other Xs on Y. **Regression model**:

Y = f (Dummy whether extension was received, level of education, farm size, …)

Ex-post facto design (8)

Then extension may be found not to be a statistically significant determinant of adoption at all, but only education, although the previous chi-square test of differences between the treatment and the control group was significant. Why? Because of positive *correlation* between extension and education: Maybe those who are better educated have received extension more frequently.

Ex-post facto design (9)

IF we take random samples, and IF our conceptual framework/theory and regression model tests for the major potential determinants of adoption (other than extension), a regression model based on an ex-post facto design can provide some support for rejecting or accepting the hypothesis that extension affects adoption of improved wheat seed.

Ex-post facto designs are weaker than experiments or quasi-experiments, but they are often the only design possible for causal social science research.

Ex-post facto design (10)

Ex-post facto research designs **view the life event variables as the treatment variables** (observed behavior of people is a function of many experiments experienced by the people themselves). The researcher has no influence on these ‚natural' experiments but the researcher can recall these experiences (through interviewing and other information sources) and relate these experiences to outcome variables of interest (the Ys).

Ex-post facto design (11)

Good ex-post facto research design requires:

- Sound **conceptual framework** that is based on theory
- **Random sampling**
- **Regression analysis** to make causal inferences or predictions (In contrast, experimental designs may not need regression analysis at all for causal inference).

Identification of the population and the sample

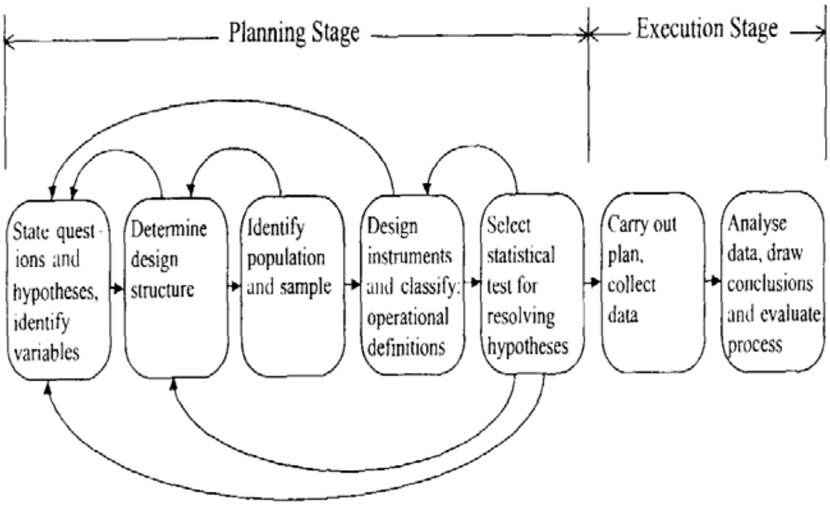


FIGURE 2.1
Stages of designing and carrying out a study, including iterations for modifications and improvements during planning

Source: Black, 1999: 27

Why take samples? (1)

- ➔ Correlational and causal research designs alike seek to draw conclusions from a sample of the general population of interest.
- ➔ How is X related to Y? Is there a causal relationship between X and Y?

Why take samples? (2)

Major reasons for taking samples:

1. Samples are taken to **represent the population** (It is too costly to study every member of the population).
2. To **control for extraneous variables** that may have an influence on Y (other than the principal Xs of interest to the researcher). There are 3 types of extraneous variables:

Why take samples? (3)

- Other variables (not of immediate interest) that the researcher identified in the conceptual framework of being a potential cause on Y.
- Unanticipated variables (those not appearing in the conceptual framework) → **always assume that they exist!**
- Variables influenced by the study itself (the type of survey, etc.). The latter may not be controlled for simply by sampling, but also by appropriate design of questionnaire and introduction of the survey to the sample. (ex. Prog impact → do not ask directly the impact of a program because of respondent bias).

Why take samples? (4)

3. Samples may often be **more accurate** than censuses (too many interviewers, too much time may decrease the quality of data).

Terms in sampling (1)

- ***Element (or observation unit)*****:** This is the unit about which information is sought. In our example, an element would be a wheat farmer.
- ***Sampling unit*****:** This is the element or elements available for selection at some stage in the sampling process. This term is only relevant for certain sampling procedures.
- ***Sampling frame*****:** This is the list of sampling units available for selection.

Terms in sampling (2)

- ***Population (or universe)*****:** This is the aggregate of all the elements defined prior to selection of the sample. A properly designated population must be defined in terms of (possible definitions in parentheses):

(1) **Elements** (wheat farmers).

(2) **Sampling units** (Farmers first, then wheat farmers being defined as farmers that planted wheat in growing season 2005/2006).

(3) **Extent** (South Nyanza in Kenya).

(4) **Time** (Survey at the end of the growing season 2005/2006).

The sampling process

Step 1: Define the population, sampling units, extent and time.

Step 2: Get a research permit if this is required in the country you work in.

Step 3: Construct the sampling frame.

Step 4: Determine the sample size.

Step 5: Select a sampling procedure.

Step 6: Select the sample.

Types of sampling procedures (1)

Two major types:

1. **Non-probability sampling**

   Useful for case study/exploratory research, but not for causal research or for making statistically valid inferences about the population.

   Several procedures exist for selecting samples, such as ***snowball sampling*** (previous respondent selects next one), ***purposive sampling*** (researcher selects respondents), ***quota sampling*** (e.g., 50% of respondents must be male and 50% female, different age groups must be proportionately represented, etc.)

Types of sampling procedures (2)

2. **Probability sampling**. There are four types:

- Simple random sampling
- Stratified random sampling
- Cluster sampling
- Stage sampling

Simple random sampling (1)

➔ There is only **one type of sampling unit** (for example all farmers living in two villages).

➔ One needs a complete list of population members, for example to be obtained from secondary data of the last village census or, if outdated, from a new census solely done for the purpose of the research.

➔ **Golden rule: Each and every element of the population has to have the same probability of being in the sample!**

Simple random sampling (2)

Methods to draw a random sample:

- Mix equal-size papers with household names, and draw them, say from a hat (or let the village head draw them) (NOT really recommended if one cannot cut equal size paper or cannot mix diligently)
- Using random selection features in EXCEL or SPSS (recommended with large population to draw from).

Simple random sampling (3)

- Write all members on a list, start from a randomly chosen point in the list, and select the sample in equal distance from the previous element chosen, going through the WHOLE list.

  For example, suppose the evaluator needs a sample of 10 from a list of 150 households. First a number is randomly selected between 1 and 15 (150 divided by 10), and starting from the ?th household on the list, every 15th one is selected. If 5 were the randomly selected number, then the sample would be composed of households 5, 20, 35, 50, 65, 80, 95, 110, 125, 140 on the list.

Stratified random sampling (1)

➔ There are **at least two types of sampling units** (for example certain types of farmers, wheat versus rice farmers)

➔ Like in simple random sampling, one needs a complete list of population members that are differentiated into the different strata. For example: Wheat-only farmers, rice-only farmers, farmers doing both crops, and farmers neither planting rice nor wheat.

Stratified random sampling (2)

➔ **Advantages:**

- Being able to ensure that specific groups (strata) are included proportionately in the sample (like their proportions in the population are),
- Being able to randomly select more of a certain stratum that is not very frequent in the population (we may, for example, be primarily interested in farmers who grow wheat, but only 12 % of farmers in a village may do so).

How to weight disproportionate samples in stratified sampling (1)

- For stratified random sampling one needs to determine the sample size in each stratum.
- The methods to draw random samples within these strata are the same as for simple random sampling.
- If the sample size in the stratum is NOT proportionate to the relative frequency of the stratum in the population, one must weight all results (SPSS has such a function called WEIGHT BY ....).

How to weight disproportionate samples in stratified sampling (2)

Formula for weighting: $W_i = (n_i / N) / (s_i / S)$

Where

W_i = Weight assigned to each element of stratum i

n_i = number of elements in stratum i

N = total number of elements in the population

s_i = number of elements belonging to stratum i in the sample

S = total size of the sample

Example: n_1=10, N=100, s_1=4, S=8.

The weight for each sampled element in stratum 1 would be (10/100)/(4/8) = 0.2.

How to weight disproportionate samples in stratified sampling (3)

It must always hold that the sum of sampling weights for all the elements selected in the sample adds up to the sample size (i.e., 8 in the example). In a simple random sample, we may think of each sampled element having a weight of 1.

The weight of stratum 1 is 0.2 as it is overrepresented in the sample (its share is 50 % in the sample, but only 10 % in the population).

If stratum 2 comprises 90 elements, they must have the weight: (90/100)/(4/8) = 1.8.

Thus, the sum of weights for all 8 elements in the sample is (4*0.2 + 4 *1.8) = 8.

Cluster sampling

➔ **Random samples of successive clusters of subjects** are chosen, e.g. bank branches, then certain credit groups, then individual members of credit groups.

➔ **Advantage**: Cluster sampling makes random selection feasible even if no single list of population members exists, but local lists do (e.g. at the bank branch level, a list of the credit groups in that branch exists, and at the group level, the leader of the group can provide a list of members).

Stage sampling (1)

➔ Stage sampling is a **combination of cluster sampling and stratified sampling**. For example, villages are grouped into two strata. One stratum is close to the road (say, less than 2 kilometers away), and one is far away. In each stratum, a random selection of villages is taken. In each of these villages, all bank branches get visited to implement a cluster sampling approach to randomly select members of credit groups.

Stage sampling (2)

➔ **Advantage**: It combines the advantages of cluster and strata sampling. It is often used.

➔ **Disadvantage**: It is a complex undertaking. The strata and clusters must be carefully defined. One needs to calculate weighted population means, and also use weights in regression analysis.

Practical methods to construct a sampling frame (Method 1) (1)

Method 1: Using published secondary data

1. **Search for secondary published data** that reports on our sampling units (names of villages or names of households). Such data are available at Statistical Offices or ministries (for specialized data on health or rural infrastructure).

Practical methods to construct a sampling frame (Method 1) (2)

2. **Judge the quality of the data**.

- When was the census conducted? Has there been much fluctuation in my sampling units since the census was done? Is it regarded to be fairly accurate by national experts?
- If we cannot answer these questions, we may do a census ourselves in a very small village, and compare the list of households of our census with the official numbers.
- For information on villages as sampling units, data published by ministries and by their regional offices are usually accurate. Data on households residing in a certain village are almost always questionable. Here, decentralized information gathering at the local level is often better.

Practical methods to construct a sampling frame (Method 1) (3)

- In most countries, the census information on households is updated by the mayor of the village regularly. Data available at decentralized levels is often more accurate than at central levels.

➔ so often obtaining a list of households from the mayor or traditional village head is better than searching through piles of data in some ministry. We may update this list with the help of the mayor‘s/head‘s knowledge. If the village is too large (for example more than 300 households), we may randomly choose certain sections of the village, e.g., administrative sub-units.

Constructing a sampling frame (Method 2) (1)

Method 2: Doing our own census

- If no recorded data on paper on the residents in a village/our research site is available, even at the local level, the only way out is to do a census ourselves. This may apply to refugee villages, or to semi-urban and urban areas with very transient populations.

Constructing a sampling frame (Method 2) (2)

- There are principally **two sub-methods**:

I. The first is to invite the village elders, teachers, religious leaders (all people that supposedly might know the residents of the village), and ask them for the names of the heads of the households living in the village. Ask them how they judge the quality of the information provided.

II. If the validity of the information is questionable, the only way out is to undertake a census by walking from house to house, and asking the people. This of course is a somewhat obtrusive undertaking, and needs to be done with the permission and active support of the village leaders.

Constructing a sampling frame (Method 2) (3)

- For certain sub-groups of households for which we may wish to stratify, we can also contact other sources of information. For example, for households with school-going children, ask the teacher. For credit group members, ask the bank or NGO. For local vendors of rice in a district, ask the two or three wholesalers of rice in the district (The latter names and their residences we may find out by going to the district government).

Random sampling without a sampling frame (1)

- If methods 1 and 2 are too time-consuming, and if we are not really interested in stratifying the population into several groups and to later estimate the mean of the total population, we may decide to do sampling without a sampling frame.

Random sampling without a sampling frame (2)

- There are various methods described in social science research books, but the one often used in field research in urban and rural areas of developing countries is the one developed by UNICEF/WHO for sampling households with pre-school children (may be termed the ***Random Walk Method***)

Source: See on this method: Henry, C., M. Sharma, C. Lapenu, and M. Zeller. 2000. Assessing the relative poverty of microfinance clients: A CGAP operational tool. Consultative Group to Assist the Poorest (CGAP), The World Bank, Washington, D.C. (http://www.cgap.org/publications/other)

Sampling without a frame: The random walk method (1)

1. Approximate the village or locality boundaries with the help of the mayor/village head (draw a map).
2. Determine a central point and assess density of households.
3. Divide area into quarters (reflecting approximately four quarters of the population).
4. Randomly select one or more directions by spinning a pen or bottle to determine the one or two quarters to be sampled (it is recommended to sample at least two quarters).

Sampling without a frame:The random walk method (2)

5. Follow selected direction and select households in intervals of a pre-selected number that allows us to do a random walk through the whole quarter and select the desired sample size from all parts of the quarter. At pre-specified intervals, we may spin the pen again to change the direction within one quarter (or if we end up in a place where we cannot continue in the initial direction).
6. Replace non-response/drop-out households by sampling the neighboring household.

Determining the sample size (1)

- The sample size required to obtain an estimate of the population mean of a certain variable X at a certain level of precision can be calculated if one *a priori* knows the standard deviation (s) of X in the population; from our sample mean of X we can derive a ***confidence interval*** within which the population mean will fall with a certain probability (typically 95%). The sample size (N) determines the width of the confidence interval, i.e., the precision of our prediction of the population mean.
- The lower and upper bounds of the 95% confidence interval can be calculated as follows:

Determining the sample size (2)

Lower bound: $X_{Lower} = \overline{X} - 1.96 \cdot s_{\overline{X}}$

Upper bound: $X_{Upper} = \overline{X} + 1.96 \cdot s_{\overline{X}}$

Where $s_{\overline{X}} = \frac{s}{\sqrt{N}}$

= Standard error of X = Standard deviation of sample means of X

- Note that +/- 1.96 are the Z-scores between which 95% of observations fall in a ***Standard Normal Distribution*** (mean = 0, s = 1). Probabilities for Z-scores can be found in statistical tables.
- Any normally distributed variable X can be converted into Z by subtracting the mean and dividing by s:

$$Z = \frac{X - \overline{X}}{s}$$

Determining the sample size (3)

- Rearranging this formula and replacing Z by the specific values below which 2.5% and 97.5% of observations lie in a Standard Normal Distribution (i.e., 95% of observations lie between them) leads to the equations for the calculation of the confidence interval above.
- Since we are interested in the variability of sample means and not the variability of observations in the sample, we have to replace the standard deviation of X by the standard error of X.
- We can now determine the sample size (N) required to derive a 95% confidence interval of a pre-specified width (W) as follows:

$1.96 \cdot \frac{s}{\sqrt{N}} = \frac{W}{2}$ Hence, through rearranging: $\left(\frac{1.96 \cdot 2 \cdot s}{W}\right)^2 = N$

Determining the sample size (4)

Example: If we knew that the standard deviation of income in our population was 100 US$, and our goal was to estimate mean income at a level of precision of 40 US$ (= W), then the required sample size would be:

$$\left(\frac{1.96 \cdot 2 \cdot 100}{40}\right)^2 = (9.8)^2 = 96.04$$

➔ If we wanted to halve the confidence interval of our estimation to 20 US$ (= W), the sample size necessary would be:

$$\left(\frac{1.96 \cdot 2 \cdot 100}{20}\right)^2 = (19.6)^2 = 384.16$$

➔ Note that there is a **quadratic relationship** between the precision of our estimate and the required sample size, i.e., to double precision (= to halve the width of the confidence interval) we have to increase N by the factor 4!

Determining the sample size (5)

- Unfortunately, we **usually do not know the distributions** of the Xs and Ys before the research. Exceptions are data published in national/prior surveys from which one can make some inferences and best guesses of what the distribution may look like.

➔ The optimal sample size is almost never being calculated and used as a decision criterion for research surveys (Exception: National-level household surveys, Demographic Health Surveys of WHO, Mother and Child surveys of UNICEF that often have sample sizes between 6,000 to 12,000, and in large countries, up to 100,000 sample households).

Some practical hints for adequate sample size (1)

Warning: *Most of the hints below we do not find in statistics and social science research books. Rather, they emanate from experience in doing quantitative social science research in rural development.*

1. The larger the sample size, the more likely it is that we find significant differences between groups. However, the gains in precision decrease quickly at the margin with increasing sample size (quadratic relationship!).

Some practical hints for adequate sample size (2)

2. The sample size decision is in practice rarely driven by the desired statistical precision, but mainly by the size of the research budget, the available time until reporting the results, and the type of research question (descriptive/correlational studies require smaller samples than causal research designs, and univariate < bivariate < multivariate analyses).

Some practical hints for adequate sample size (3)

3. Time: If we spend more than 40 % of our research time on data collection, we need to have good justifications for doing so (may be okay for exploratory, case-study research but not for quantitative research).

Some practical hints for adequate sample size (4)

4. The larger the sample size, the better. BUT: This rule only applies if the quality of the survey instrument and all other phases of research, including data cleaning and analysis, do not suffer because we spend most of our time, motivation and money collecting data. Usually, with a larger sample size, so-called non-sampling errors increase, such as interviewer errors, non-response errors, and data processing errors.

➔ **Note that collecting data is NOT research unless they are properly analyzed and reported in an essay or thesis!**

Some practical hints for adequate sample size (5)

5. Descriptive studies that seek to document frequencies, means, and bivariate cross-tabulations may yield reasonable results with a sample size between 60 to 100. Univariate analyses (i.e., exploring the distribution moments of one variable) require smaller sample size than bivariate analyses (i.e., exploring the relations between two variables).

Some practical hints for adequate sample size (6)

6. Causal studies (usually multivariate analyses with many Xs and one or few Ys), which employ simple ordinary least squares (OLS), Probit and Logit regression models, should have a minimum sample size of 100.

Some practical hints for adequate sample size (7)

7. Causal studies that wish to employ complex regression, such as two-stage and three-stage regression models, should have sample sizes above 200.
8. A lower sample size is sufficient if our questionnaire is carefully pre-tested so that interviewer and respondent errors are low (which reduces the standard deviation of variables).

Some practical hints for adequate sample size (8)

9. A smaller sample size is sufficient for the same precision if we employ stratified random sampling instead of simple random sampling.
10. Holding budget and time constant, we can increase our sample size by reducing the length of our questionnaire (i.e. the scope of our research), thus spending less time per respondent. However, there is a considerable fixed cost in sampling each respondent, and in visiting them.

Design instruments and classify operational definitions (1)

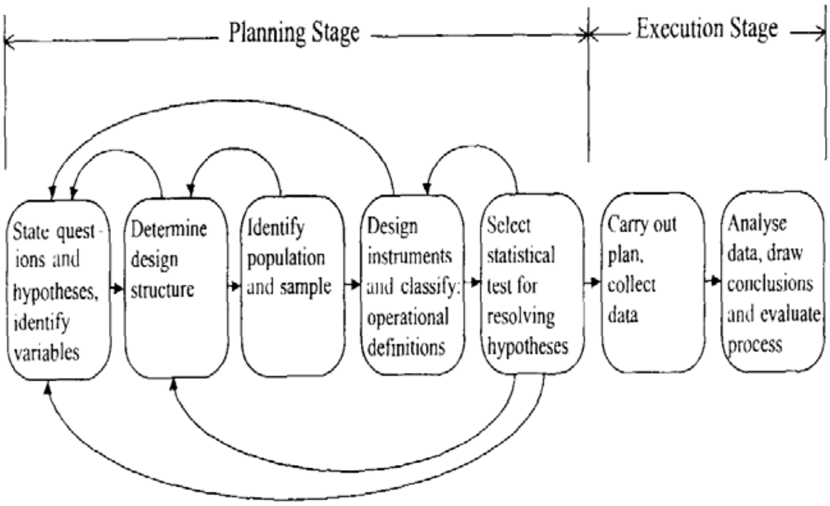


FIGURE 2.1
Stages of designing and carrying out a study, including iterations for modifications and improvements during planning

Source: Black, 1999: 27

Design instruments and classify operational definitions (2)

1. Identifying empirically operational measures for the variables in our conceptual framework. These are: The Ys, the Xs (whether observed or, if we do an experiment, manipulated) This mainly requires choosing the measurement scale of the variables (whether nominal, ordinal, or interval/ratio).

Design instruments and classify operational definitions (3)

2. Designing the questionnaire:
- Deciding on the structure of the questionnaire (which questions first, which later)
- Layout of questionnaire (for easy training of enumerators and easy data entry later)
- Phrasing of questions

Design instruments and classify operational definitions (4)

3. Pre-testing the questionnaire (in near-field conditions), e.g. in a neighboring village that has not been randomly selected by us.

4. Revising/finalizing the questionnaire.

Measurement scales for variables (nominal variables) (1)

1. Nominal variables: Unique definition of numerals, e.g. 1, 2, ... 9.

→ These are often **categories**, such as region, religion, ethnic group, occupation, etc.

Measurement scales for variables (nominal variables) (2)

Note: It is good practice not to enter so-called string variables, e.g. names of villages, into the spreadsheet if we wish to analyze the data. We always re-define them into nominal variables. The more detailed the code (the more numerous the categories), the better. However, it also does not make sense to distinguish categories that do not really appear in practice. On the other hand, we try to avoid the famous ‚other', or the even worse one ‚none of the above'. Cases falling into these categories are data with *no information*.

Measurement scales for variables (nominal variables) (3)

Permissible statistical procedures:
Percentages, mode, Chi-Square test

Examples:
1= yes, 2 = no
1 = Male, 2 = Female
District: 1 = Kasungu, 2 = Liwonde 3 = Lilongwe

Measurement scales for variables (ordinal variables) (1)

2. Ordinal variables: Order of numerals in a logical/hierarchical sense, e.g. 1 < 2 < 3

→These are often variables measuring attitudes, preferences, levels of achievement (education, salary ranks, farm size ranks, etc.)

Measurement scales for variables (ordinal variables) (2)

It is good practice to **carefully pretest** prespecified codes for ordinal variables. Often, we forget some possible combinations (This can jeopardize the validity of the variable for analysis).

Permissible statistical procedures: All tests permissible for nominal variables, plus percentiles, median, and rank-order correlation (Spearman‘s corr. coeff., Kendall‘s Tau).

1. Quality of last harvest (subjective rating by farmer):
1= below average, 2 = about average, 3= above average
2. Education
1= did not complete primary school, 2 = did not complete secondary school, 3 = secondary school education or higher

Measurement scales for variables (interval variables) (1)

3. Interval variables: Equality of differences between levels of measurement (2 -1 = 7-6).

➔These are often variables measuring attitudes, opinions, index numbers)

Note: The difference between ordinal and interval data is that the difference between any two ranks is equal to the difference of any other two ranks.

➔In order to use this as true interval data, the researcher should phrase the question with the precursor ‚*On a scale from 1 to n, how would you ...?‘*

Measurement scales for variables (interval variables) (2)

Permissible statistical procedures: All tests permissible for nominal and ordinal variables, plus range, mean, standard deviation, and product-moment correlation (Pearson‘s correlation coefficient).

On a scale from 1 (= very poor) to 5 (= excellent), how would you rate the quality of the wheat extension program in your area? The possible answers are any number between 1.0 and 5.0.

Measurement scales for variables (ratio variables) (1)

4. **Ratio variables**: A ratio variable has all the qualities of an interval variable plus a zero point.

→ These are variables measuring ages, distance, weight, number of visits, sales, income, etc.

→ We can calculate ratios, e.g., person A weighs twice as much as person B.

Measurement scales for variables (ratio variables) (2)

Note: It is important to use **units of measurement** that are known and normally used by the respondent. E.g., if the respondents measure the quantity of rice harvested in bags, then we ask them how many bags they harvested. We do not ask about kilograms or tons. Before the survey, we may go to the village/ local market, and weigh a bag of rice. However, if there is no freshly harvested rice around, we may ask the extension officers how much one bag weighs in kg.

Measurement scales for variables (ratio variables) (3)

Permissible statistical procedures: All tests permissible for nominal, ordinal and interval variables, plus geometric and harmonic mean.

How many kilograms of rice did you and your family members eat yesterday and the day before?
Note: In this example, people buy and measure rice in kilograms.

Best practices in questionnaire design (1)

A questionnaire is a standardized form for collecting data from respondents for the purpose of measurement.

It contains questions that are to be asked **in the same way to all respondents**, with answers usually being recorded as numbers using standardized sets of response categories.

Questionnaire design is more of an art than a science.

Best practices in questionnaire design (2)

The quality of the questionnaire depends on skill and judgement, a clear concept of what information is needed, how the data will be analyzed, and an awareness of possible sensitivities of respondents.

Good questionnaires are often developed in stages and involve local researchers and extensive pretesting.

Best practices in questionnaire design (3)

The phrasing of a question determines whether the researcher and the respondent interpret the meaning of the question in the same way.

No single phrasing of a question is correct. Instead it is important to understand clearly what effect a particular wording can have on the response

Phrasing of questions (1)

The words used in a question should be familiar to the interviewer and respondent. They should correspond to local word usage and practices (if possible, use specific terms of local word usage even in your English/French/Spanish/etc. version). If the questionnaire is administered in a local language that we do not command, the training of the enumerators, and the development of a common understanding and knowledge how to phrase the questions in local language becomes extremely important. The final result of this training can be (but must not be) a questionnaire translated into local language.

Phrasing of questions (2)

The questions should not be ambiguous (with more than one meaning possible).

For example: *How high is your income?* This is a sensitive and ambiguous question.

The respondent might ask himself: “From which type of occupation, during which period, for me alone or for the family members, is the interviewer asking gross revenue or revenue minus variable costs. This is too complicated for me, let’s just give her or him a number. Or the respondent asks the interviewer: “What do you mean?”

Phrasing of questions (3)

Well-posed questions:

1) Use simple and clear words

2) Avoid biasing the answer,

No: Was the harvest good last year?

Yes: How was the harvest last year?

3) Avoid implicit alternatives and assumptions,

No: “Can you afford a telephone” carries the assumption that the respondent cannot afford it.

Yes: “Do you have a telephone in your house for your own use?”.

Phrasing of questions (4)

4) Avoid estimates (if possible). However, many questions make estimates inevitable (e.g. yield, harvest, amount consumed)

No: *We do not ask amount of rice consumed in last year, if amount of rice consumed in last week could suffice for analysis.*

Yes: *We try to make the estimate as simple as possible without jeopardizing too much our research objectives.*

Phrasing of questions (5)

5) Avoid double-barreled questions

No: *If you are not a farmer, do you earn your living from trade? And if you are neither a farmer nor a trader, are you earning your income from …?*

Yes: *What is your main source of income?*

6) Consider the frame of reference

No: *Was any cattle stolen?*

Yes: *During the last 12 months, was any of your cattle stolen?*

Phrasing of questions (6)

7) Do not ask questions about third parties

No: *Do households in your village sell wheat in the market?*

Yes: *Do you sell wheat in the market?*

Golden rule: Always try to ask the person directly to whom the question applies (people do not like to be informants about the behavior of others, or may not know such behavior).

Phrasing of questions (7)

8) Avoid asking hypothetical questions

No: *Would you take a loan if the interest rate was only 10 percent?*

It is a leading question (word *only* invites to say yes), but even if we drop the ‚only', most people may say yes anyway because this is not a real situation but only one presented by us. So, why not take a loan from somebody whom I do not know, but the interviewer seems to know? Such questions may elicit dreams, wishes, or things that the respondent thinks the interviewer wants to hear, but have limits in telling about actual behavior of people. However, sometimes these questions can be useful (especially if you cannot recall past behavior for the research questions of interest). In this case, put them at the end of the questionnaire.

Three principal types of questions (1)

1) **Open-ended questions**

Examples:

1.1 *In your opinion, what weaknesses does the agricultural extension service for wheat have?*

This allows for a detailed answer. However, interviewer may have to write a lot (forget a lot). → problem of interviewer bias. Open-ended questions are good for qualitative answers, especially for exploratory research in small samples.

1.2 *How old are you?*

1.3 *What school education do you have?*

Three principal types of questions (2)

Examples 1.2 and 1.3 are open-ended questions for which answers are clearly defined and pre-known. In case of 1.2, the age in years is recorded, in case of 1.3 a pretested code can be used so that the interviewer after the interview can correctly categorize the response.

Open-ended (but pre-coded) questions are the most preferred as they influence responses less than multiple-choice or dichotomous questions.

Three principal types of questions (3)

2) **Multiple-choice question** (in a personal interview, ***the interviewer reads the answers***)

2.1 *Please indicate the two most important constraints in the ag extension service for wheat among the following constraints: 1= does not visit frequently 2 = officer does not know much 3 = too expensive 4 = not relevant for my farm*

2.2 *Give age groups: 1= Under 18; 2 = 18-55; 3 = above 55*

2.3 *Give categories of school education: 1= no school 2= primary school 3 = above primary*

Three principal types of questions (4)

3) **Dichotomous question**

Are you satisfied with the wheat extension service? 1= yes, 2 = no, 3 = no opinion**.**

Questionnaire design: Coding of questions (1)

Each and every question in a questionnaire either asks for nominal, ordinal, or interval/ratio data. A first test whether the question is a potentially good one is to ask ourselves the **units**. For interval/ratio data, the units are often monetary, weight, distance, or other straightforward measures.

For nominal (categorical) and ordinal data, **potential answers may be pre-coded**.

Questionnaire design: Coding of questions (2)

The coding of the questionnaire for this type of variables is very important, unless we want a verbal answer that we do code later for quantitative analysis. However, for larger sample surveys that are not done personally by us, but by interviewers, the questionnaire should already contain pre-coded answers.

Questionnaire design: Coding of questions (3)

The codes should be listed on the same page of the questionnaire (preferred), or in a separate interviewer guide.

Later, after data entry, we can test whether any data point violates the pre-specified codes

→ one step of data cleaning.

Structure of questionnaire (1)

Structure of questionnaire/Sequence of questions:

- Questions should **logically flow**. Do not jump from household composition to income, and back to education of hh members, then to land possession, and then back of whether the children have attended school.
- A questionnaire should be structured into **different modules**, each covering a certain topic. The less sensitive topics should be placed at the beginning (usually we begin with the demographic info on the members of the households). The most important and most difficult questions should be raised about 15 to 20 minutes into the interview.

Structure of questionnaire (2)

- Are the questions sensitive, i.e., may the question offend the respondent? If yes, we try to avoid the question. If it cannot be avoided because it is REALLY, REALLY important, we put it at the end of the questionnaire, and ask it in the last module.
- Questions regarding attitudes, preferences, subjective ratings of other people or institutions should come last. They are almost always sensitive.

Structure of questionnaire (3)

A questionnaire usually contains the following five sections:

(1) **Identification data:** For example name of household and village. This information should appear on each and every page of the questionnaire, preferably as a header.

Structure of questionnaire (4)

(2) Request for cooperation: This is an optional question. We may include a couple of sentences on the first page of the questionnaire that the interviewer is asked to read before the survey commences in order to request the cooperation of the respondent. The cooperation request should always state that the survey will not have any direct benefits (such as a new project/service) for the respondent, and that all information is treated anonymously and not being disclosed to anybody within or outside the village.

Structure of questionnaire (5)

(3) **Instructions for interviewers:** For certain questions, we may add instructions to the questionnaire right above or below the question. However, most of the instructions need to be orally discussed and agreed upon in the enumerator training, or be put on a written guide separate from the questionnaire.

(4) **Information section:** This is the main section of the questionnaire. It is usually structured into different modules that logically flow and capture different topics.

Structure of questionnaire (6)

(5) Classification data section: This information obtains the socio-economic characteristics of the respondent and his/her household and/or village. In practice, a questionnaire at household level usually has a table that lists all the household members, and then asks about the demographic and socio-economic information. Very often, survey questionnaires do not contain enough classification data. These data often generate useful independent variables for future analysis. We always carefully check whether the classification data are sufficient for our research questions.

Measurement of attitudes

Attitudes = beliefs, perceptions, personal evaluations and likes/dislikes

Attitudes are generally considered to have three main components:

(1) A cognitive component - a person's awareness and knowledge about the object of concern

(2) An affective component – a person's feelings about the object, such as "good" or "bad"

(3) A behavioral component – a person's readiness to respond behaviorally to the object.

Attitudes influence social and economic behavior, and are therefore important in most social science research.

Scales in the measurement of attitudes (1)

Important scales in structured questionnaire administered by an interviewer are:

1) Verbal ranking scales

Example: I find the agricultural extension service

1 = very good
2 = good
3 = acceptable
4 = bad
5 = very bad

(Ordinal variable)

Scales in the measurement of attitudes (2)

Issues in verbal rating scales:

- number of categories (not more than five or six)
- Odd or even number of categories (no recommendation, but odd has a neutral value)
- Balanced versus unbalanced scale (more positive than negative answers may bias results, but can be appropriate if pre-test found that there are more positive than negative responses with a balanced pre-test question)

Scales in the measurement of attitudes (3)

- Extent of verbal description (should be short and precise)
- Category numbering: In principle, most attitude measurements result in ordinal variables, not interval variables. Adding “On a scale from 1 (= very bad) to 5 (= very good), how would you rate…” would result in an interval variable.
- BUT: When applying questions that are based on such a scale, make sure that the respondent understands this concept!

Scales in the measurement of attitudes (4)

2) Rank-order scales

- Example: Among the following constraints in the agricultural extension service, which do you judge to be the most important, the second most important…
- After stating the question, the interviewer reads out a list of pre-coded constraints.
- This is a simple technique that is often used.
- Issue: This technique again only produces ordinal data (constraint 1 is more important than constraint 2, but by how much remains unknown).

For other scales, see for example Black, 1999

Best practices in data entry (1)

We can save much time in data analysis if we design our data entry template properly. Here are some of the most important best practices:

1. **Each case entered in an SPSS file begins with the so-called key variables.** The key variables identify the case. For example:

- Code for name of village
- Household code
- Code I.D. for a specific member of a household

Best practices in data entry (2)

2. **We never aggregate information from the questionnaire before data entry.**

Example: If we ask the age and education level for each member of the household, we do not define variables AGE1, EDUC1, AGE2, EDUC2, and so on. Instead we create a member-specific file (with the key variables as listed for example under point 1), and have only ONE variable for AGE. SPSS has very convenient functions to aggregate information to any desired level, but it is VERY cumbersome to break the information up again to lower levels.

Best practices in data entry (3)

3. **We avoid big files** (with more than 100 variables in one file). Instead, we break the data base up in several files, usually corresponding to the different modules of the questions, further differentiated by different levels of observation units (whether hh member, plot, animal, micro-enterprise, loan received, or asset).

Best practices in data entry (4)

4. There is no variable in a data file that does not have a:

- **Variable label** that says something in brief words about the variable, and, if it is a ratio/interval variable, the **unit of measurement** (kg, Rupiah).

- **Value label** if the variable is a nominal or ordinal variable. Moreover, it is GOOD practice to retain exactly the same coding system that is specified in the questionnaire (if possible).

Best practices in data entry (5)

5. SPSS, MS Access, and other data entry programs, have **checks for value ranges** (for example, if it is specified that age cannot be above 120 years, then any value above that will be rejected).
6. **Choose variable names that have some meaning**, like age, gender, hhsize, not x23 or y24.

Best practices in data cleaning (1)

Data may contain at least **three types of errors**:

1. **Data entry error**. The typist did not enter the value recorded in the questionnaire.
2. **Interviewer error**. The interviewer did not write the correct value or any value at all for the response given.
3. **Respondent error**. The respondent did not provide a "correct" value.

➔ A number of routines are available in SPSS to identify errors, and possibly to clean the data. The following practices for data cleaning are frequently used:

Best practices in data cleaning (2)

1. Check for so-called ***wild codes***. We want to clear the data set of all codes that do not exist for a particular variable (Example: The level of schooling is an ordinal variable, and has a predefined range from 1 to 5. All other values not lying in this range are wild codes). The SPSS procedure for frequencies, and the EXPLORE, function can be used. If wild codes are found, the original questionnaire(s) must be used to re-enter the correct code value. If the questionnaire contains an incorrect value, it is most likely an interviewer error. Often, the case may need to be recoded to *missing value* (SYSMIS). This means the case is not being analyzed by SPSS if the routine uses the variable with the missing value.

Best practices in data cleaning (3)

2. ***Consistency Checks***. Checks on the logical patterns of answers can also be used to find data errors. A household that indicates it has not had a shortage of food in the past 30 days, would not also have a response to how many days members had too little food in the same time period. Consistency checks can be done in several ways within SPSS. One method is to first filter the data set for only cases responding in a certain way (using an IF Statement), and then running a frequency test on a second variable to check for inconsistencies.

Best practices in data cleaning (4)

➔ For example, households indicating they had no food shortages in the past month could be selected, and then a frequency test on ‘number of days without enough food’ could be run. If the data is clean, the results of the frequency test should show only missing values. Another method for variables with only a few categories of responses is to run cross tabulations where the responses for one variable are cross-checked in tabular form against the responses for another variable.

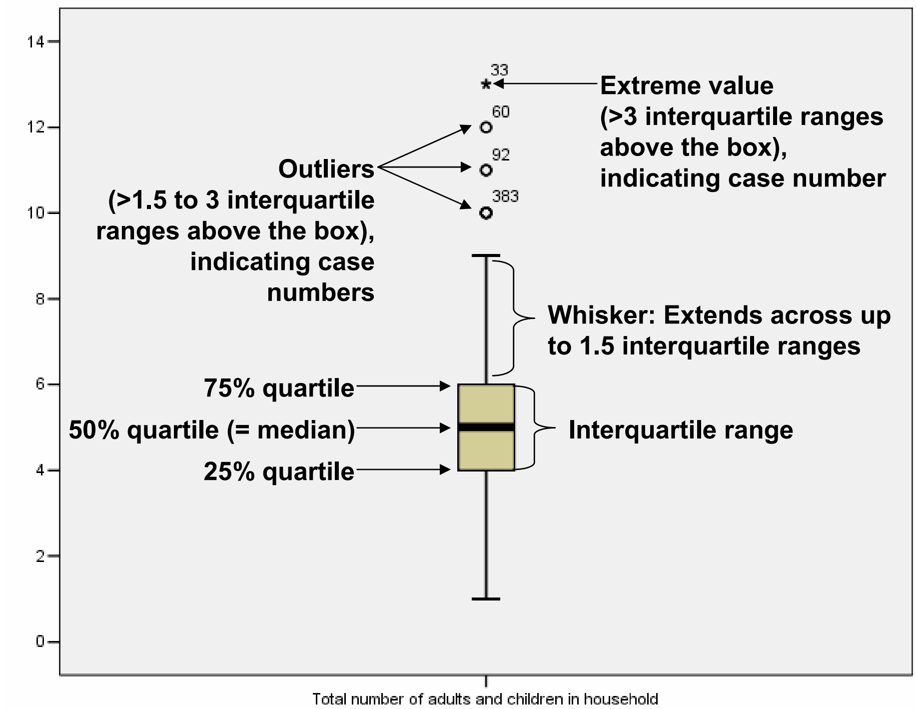
Best practices in data cleaning (5)

3. ***Extreme case check*** (so-called outliers). In some cases, responses to a question can seem highly improbable either because they are extreme when compared to the responses given by other households, or because they seem improbable given other responses from the same household. In a recent survey, a household with few assets, limited food supplies, a poor diet and low expenditures on food and clothing was found to hold land assets worth nearly $500,000. Not only was the value much higher than all other households in the survey, but it also seemed inconsistent with the household’s other responses. It was found that the data entry person had typed too many zeros in the landholding variable cell.

Best practices in data cleaning (6)

➔ Extreme cases can be identified through several techniques in SPSS. Perhaps the easiest is creating a '***boxplot***' of variable responses. Boxplots are applicable for interval/ratio variables. The boxplot shows graphically the median value of the variable, the 25% and 75% quartiles (= *interquartile range* = the 'box', i.e., the range in which the 'middle' 50% of responses fall), and it singles out all responses that lie more than 1.5 interquartile ranges above the 75% or below the 25% quartiles (*outliers* and *extreme values*), indicating the case numbers. This allows the analyst to quickly identify extreme cases.

The Boxplot



Best practices in data cleaning (7)

4. In **regression analysis**, outliers may be influential in determining the results. It is good practice to check whether outliers exist, and if so, whether they can be assumed to be true values, and therefore be retained, or whether they should be recoded to *missing value*, i.e., the case is then excluded from the regression analysis.

➔ Sometimes the analyst may recode an outlier/extreme value to the mean of the sample in order to retain the case in the regression analysis.

➔ **BUT**: This procedure decreases the standard error of the respective regression coefficient.

➔ it is only justifiable for very few extreme cases in a large sample!

Select statistical test for resolving hypotheses

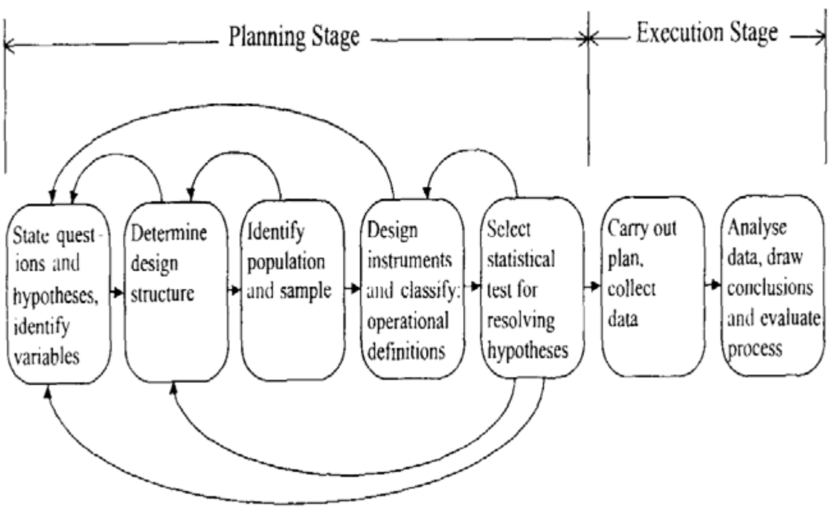
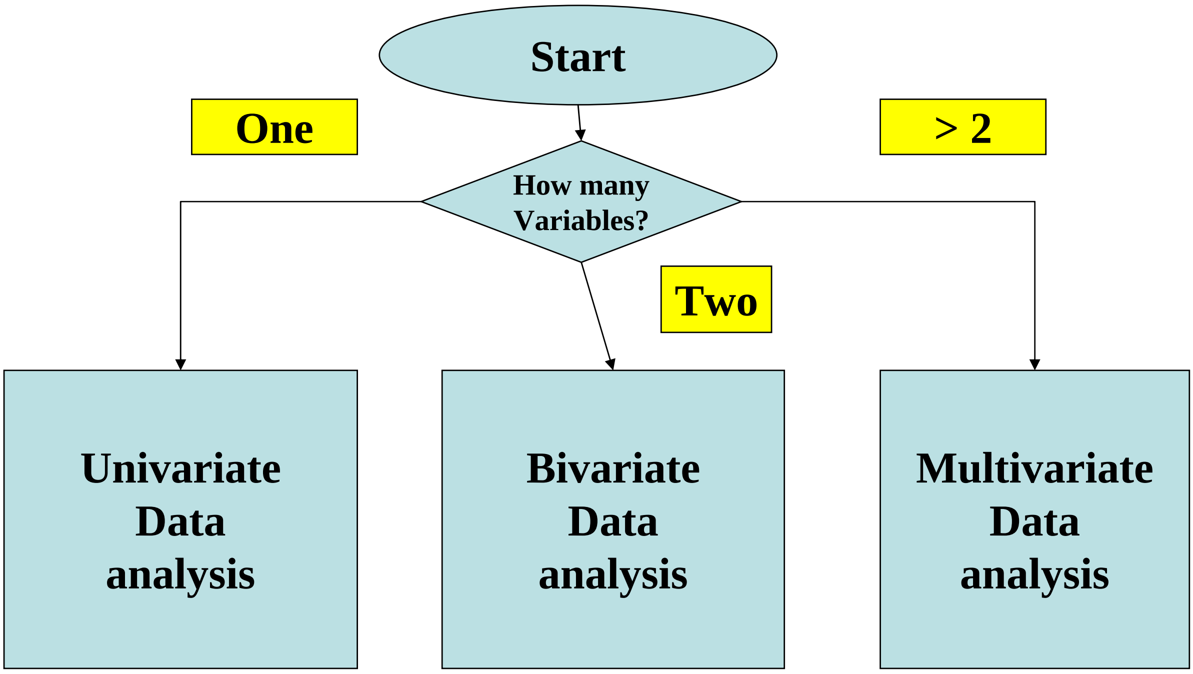


FIGURE 2.1
Stages of designing and carrying out a study, including iterations for modifications and improvements during planning

Source: Black, 1999: 27

Overview of data analysis techniques



What data analysis procedure should we use? (1)

1)How many variables are to be analyzed at the same time?

UNIVARIATE, BIVARIATE, MULTIVARIATE

2)Do we want description or inference questions to be answered?

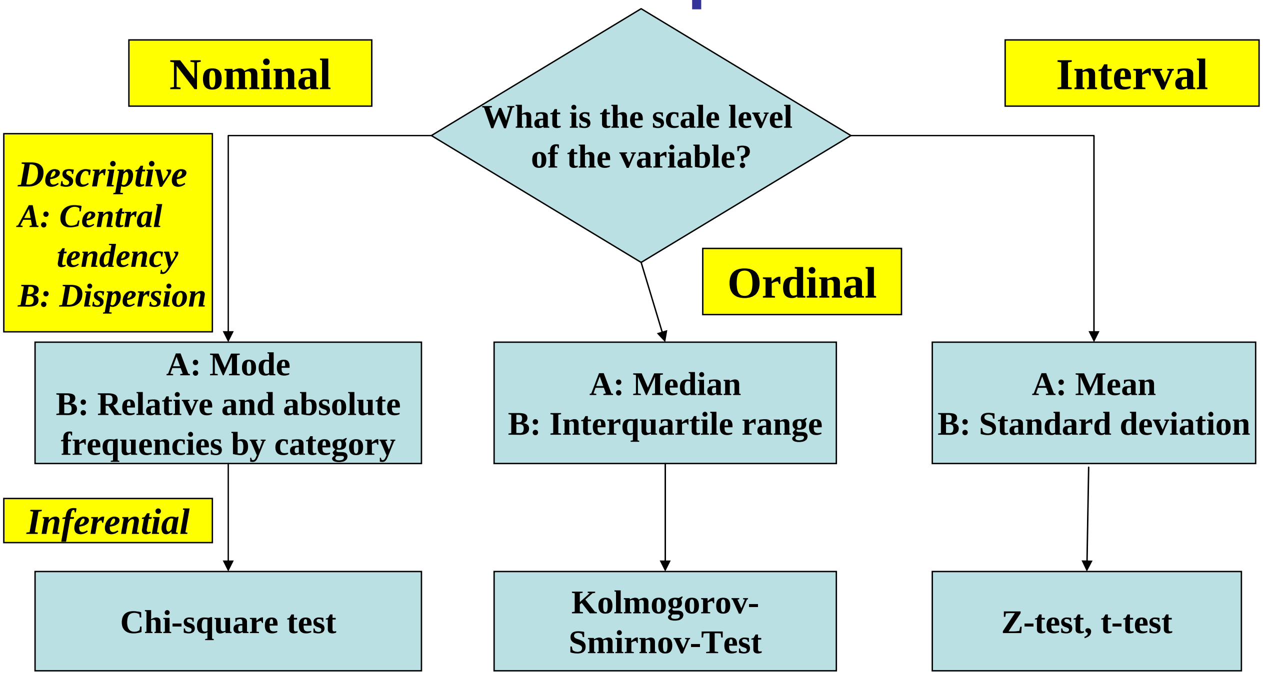
DESCRIPTION
Do we want to describe the characteristics of the sample?
Ex.: How many Christians and Muslims (frequency), average age of household heads in the sample.

INFERENCE
Do we want to make judgements about the characteristics of the whole population, not just the sample?
Ex.: Is the average age of the population 25 years?
Are levels of education different between two ethnic groups?

What data analysis procedure should we use? (2)

3) What level of measurement (nominal, ordinal, or interval/ratio) is available in the variable or variables of interest?

Overview of univariate analysis techniques



Descriptive statistics in univariate analysis: Measures of central tendency

(1) Nominal data →

Mode = the category of a nominal variable that occurs with highest frequency

(2) Ordinal data →

Median = Midpoint value of values for all cases that are ordered by value from lowest to highest

(3) Interval data →

Mean = Sum of values for all cases divided by the number of cases

Descriptive statistics in univariate analysis: Measures of dispersion

(1) Nominal data →

Relative and absolute frequencies. Example: % of literate persons in a sample, and number of literate persons in a sample

(2) Ordinal data →

Range = difference between maximum and minimum values for all cases

(3) Interval data →

Variance = Sum of the squared differences from the mean for the values of all cases. This sum is then divided by one less than the number of cases to obtain the variance. Standard deviation = Square root of variance (same unit as the measurement scale of the variable under consideration)

Inferential tests in univariate analysis (1)

(1) Nominal data: *Chi-square test*

➔ The hypothesized (a priori expected) distribution of a population is compared with the observed distribution in a sample. The default is that a priori one expects an equal distribution of all categories. The null-hypothesis is that there is no difference between the two distributions.

➔ Example: The relative frequencies for occupation in a sample are 40% farmers, 45% laborers, and 15% off-farm micro-entrepreneurs. The chi-square test determines whether there is a significant difference (at a certain level of probability of error) between the sample distribution and the expected (equal) distribution.

(2) Ordinal data: *Kolmogorov-Smirnov-test*

➔ Can test the hypothesis whether the observed distribution in a sample follows the pattern of a certain type of probability distribution (such as a normal distribution, equal or poisson distribution)

Univariate analysis: SPSS output from a Chi-Square-test

- **The Chi-Square test is useful for all nominal and ordinal data.**
- **The null-hypothesis (if not otherwise specified) is that the categories are equally distributed.**
- **Example of SPSS-Output for Chi-square test:** Self-assessment of households into 3 wealth categories (N=500):

Wealth assessment of household

	Observed N	Expected N	Residual
poor	63	166,7	-103,7
not poor	387	166,7	220,3
rich	50	166,7	-116,7
Total	500		

Test Statistics

	Wealth assessment of household
Chi-Square[a]	437,428
df	2
Asymp. Sig.	,000

Test value = 437.43 > critical value = 5.99 (from Chi-square-distribution table for P = 0.05, df = 2 = # categories - 1)

➔ Reject the null-hypothesis (i.e., equal distribution of wealth categories in the population) at P < 0.001!
(the error probability in rejecting H_0 is directly given in the SPSS output)

Inferential tests in univariate analysis (2)

(3) Interval data: *T-test*

➔ This test allows to compare the mean generated from a sample with a mean hypothesized to exist in the population, and to decide whether the hypothesized population mean is true (whether it is supported by the sample or not).

➔ Critical values of the t-distribution depend on the desired level of error probability, the sample size, and whether we want to use a one-tailed or a two-tailed test. For large samples t-values converge to z-scores. E.g., the critical t-value for an error probability of 5% (two-tailed) is 2.23 for a sample of 10, 1.98 for a sample of 100, and approx. 1.96 for a sample size of 400. Critical values are tabulated in any statistics book.

➔ Note: The t-test is a *parametric test*, which means that its application requires the data to be *normally distributed*!

Inferential tests in univariate analysis (3)

➔ You can use the Kolmogorov-Smirnov (K-S) test to check whether the distribution is normal. If it is not, you should check for outliers skewing the distribution that may need to be excluded from the analysis and, if this doesn't help, you may consider transforming the data, e.g., by taking the logarithm or square root (see statistics book on that).

➔ However, it is always useful to plot the distribution of the data since, for large samples, the K-S test easily rejects the null-hypothesis of a normal distribution, although the distribution may visually be 'close to normal'.

➔ Example: The average nitrogen application per hectare in a sample is 50 kg, but a priori information led us to formulate a hypothesized value of 55 kg. The null-hypothesis is that the nitrogen level is 55. The alternative hypothesis is that the population mean is not 55. We use a two-tailed t-test to find out whether the null-hypothesis can be rejected or not. If the t-value exceeds a certain critical level (at a given probability of error/ significance level), the null-hypothesis is rejected, and the alternative hypothesis is accepted. In other words, the sample does not support our hypothesis that farmers apply 55 kg of nitrogen per hectare.

Univariate analysis: SPSS output from a t-test

- The t-test is for all interval/ratio data that are approximately normally distributed. The null-hypothesis is that the mean in the population has a certain expected value.
- Example: We expect that the average number of cows and calves per household is 2.The sample size is 500, and the sample mean is 1.45. The SPSS output is as follows:

One-Sample Statistics

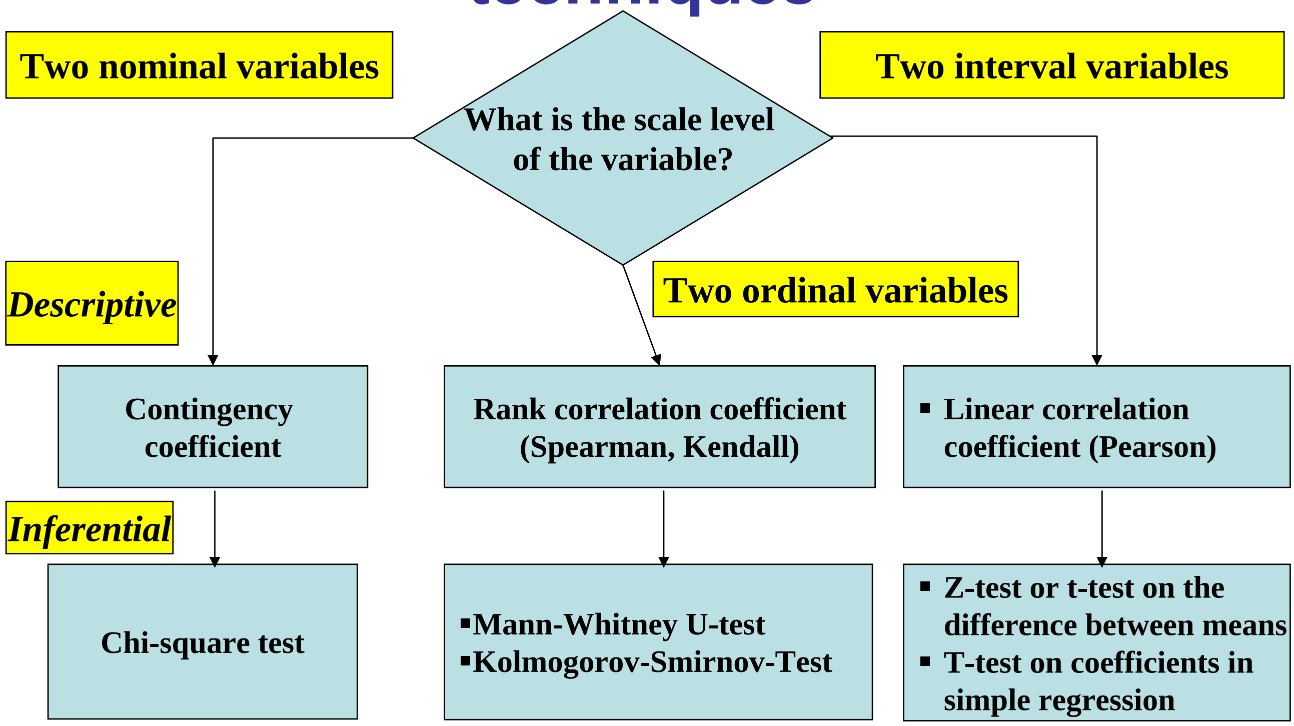
	N	Mean	Std. Deviation	Std. Error Mean
number of cows and calves	500	1,45	2,766	,124

One-Sample Test

	Test Value = 2					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
number of cows and calves	-4,446	499	,000	-,550	-,79	-,31

|Test value t| = 4.446 > critical value = 1.96 (from t-distribution table for $P = 0.05$, df = 499 = N - 1)
➔ Reject the null-hypothesis (i.e., population mean = 2) at $P < 0.001$!

Overview of bivariate analysis techniques



Descriptive statistics in bivariate analysis of interval data (1)

Pearson's correlation coefficient (r)

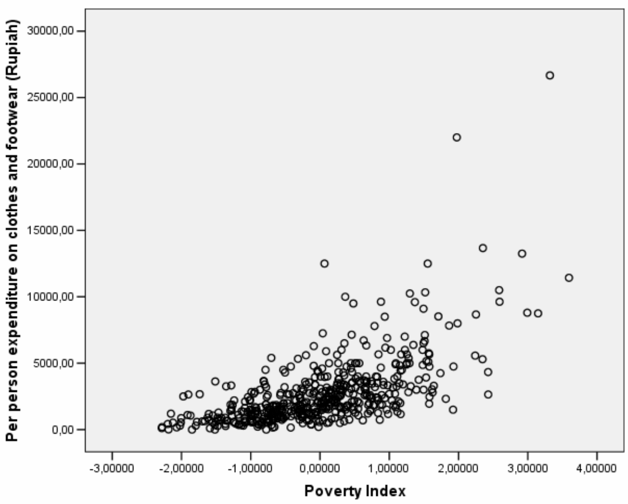
→ **r is a standardized measure of the degree of *linear association* between two interval variables.**

→ **This gives no indication of causality!**

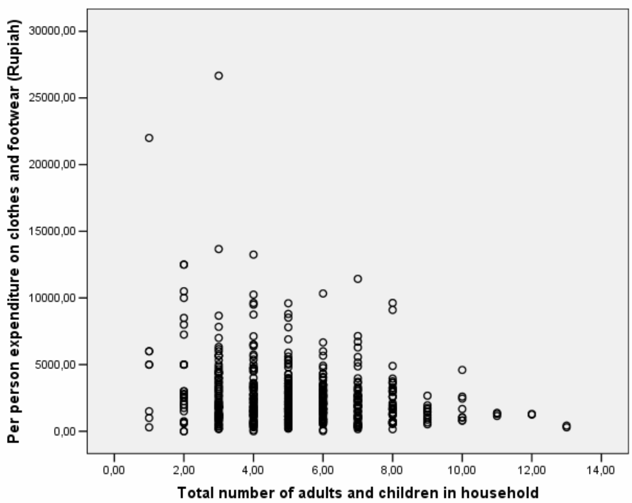
→ **If r = 0 → no association between X and Y**
r > 0 → pos. correlation
r < 0 → neg. correlation
|r| > 0.8 → very strong correlation
|r| 0.4 – 0.8 → moderate correlation
|r| < 0.4 → weak correlation

→ **r can be visualized by plotting X against Y on a scatter plot.**

SPSS-Output for correlation: Scatterplot and Pearson correlation coefficient



(r = 0.630, *P* < 0.001; N = 500)



(r = - 0.151, *P* < 0.01; N = 500)

Inferential tests in bivariate analysis (1)

(1) Nominal data: *Chi-square test*

➔ Used in so-called cross-tabulations of two variables.

➔ The chi-square test is useful for nominal and ordinal variables. It can also be applied to interval variables if we first group them into categorical information (e.g. age ➔ age group).

➔ Typical questions are:

- Is there a relationship between having heard an extension message and the adoption of a modern crop variety?
- Is there a relationship between gender and participation in a credit scheme?

➔ The null-hypothesis is that the two variables are independent from each other. This hypothesis is rejected at a certain probability of error if the chi-square value reaches a certain critical level.

SPSS Output for cross-tabs and Chi-Square-Test

The following crosstab explores whether there is a relationship between the incidence of using electricity by a household and the condition of the house itself (whether in good condition or not), as was rated by the interviewer:

Whether household uses electricity * Condition of House Crosstabulation

			Condition of House		
			Needs repairs or is delapidated	In good condition	Total
Whether household uses electricity	No electricity	Count	97	262	359
		Expected Count	77,5	281,5	359,0
		% within Condition of House	89,8%	66,8%	71,8%
	Uses electricity	Count	11	130	141
		Expected Count	30,5	110,5	141,0
		% within Condition of House	10,2%	33,2%	28,2%
Total		Count	108	392	500
		Expected Count	108,0	392,0	500,0
		% within Condition of House	100,0%	100,0%	100,0%

Chi-square test value = 22.08 > critical value = 3.84 (*P* = 0.05, df = 1 = (# rows – 1)*(# columns – 1))
➔ Reject the null-hypothesis (i.e., equal distribution) at *P* < 0.001!

Inferential tests in bivariate analysis (2)

(3) Interval data: *T-test on the difference between means*

- ➔ **It tests whether means of two sub-samples are significantly different from each other or not.**
- ➔ **Question: For example, is the average age of modern farmers lower than that of traditional farmers?**
- ➔ **The null-hypothesis is that the mean age in the two groups is not different.**
- ➔ **The critical values of the t-distribution are tabulated and can be found in any statistics book.**
- ➔ **Note that for data that are not normally distributed there are *non-parametric tests* available that are based on the *ranks* of the values rather than the values themselves (e.g., Mann-Whitney test, Wilcoxon rank-sum test).**

SPSS Output for a t-test of means in two independent samples

Example: Do members of a certain program have higher or lower clothing expenditures than non-members?

➔ **The null-hypothesis is that the mean expenditures in the two groups are not different.**

Group Statistics

	household is participating in program	N	Mean	Std. Deviation	Std. Error Mean
Per person expenditure on clothes and footwear (Rupiah)	No	300	2383,1601	2141,37516	123,63235
	Yes	200	3277,2996	3131,59104	221,43693

Independent Samples Test

$(df = n_1 + n_2 - 2)$

	Levene's Test for Equality of Variances		t-test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Equal variances assumed	7,149	,008	-3,792	498	,000	-894,13949	235,79547	-1357,42	-430,863
Equal variances not assumed			-3,526	321,603	,000	-894,13949	253,61244	-1393,09	-395,191

Simple regression (1)

➔ A *simple regression* measures how *one* independent variable (X) is related to a dependent variable Y. On the basis of the regression, predictions about the values of the dependent variable can be made based upon knowledge of the values of the independent variable. Regression is also used for *causal analysis* using quasi-experimental and ex-post research designs.

➔ General structure of a simple regression model:

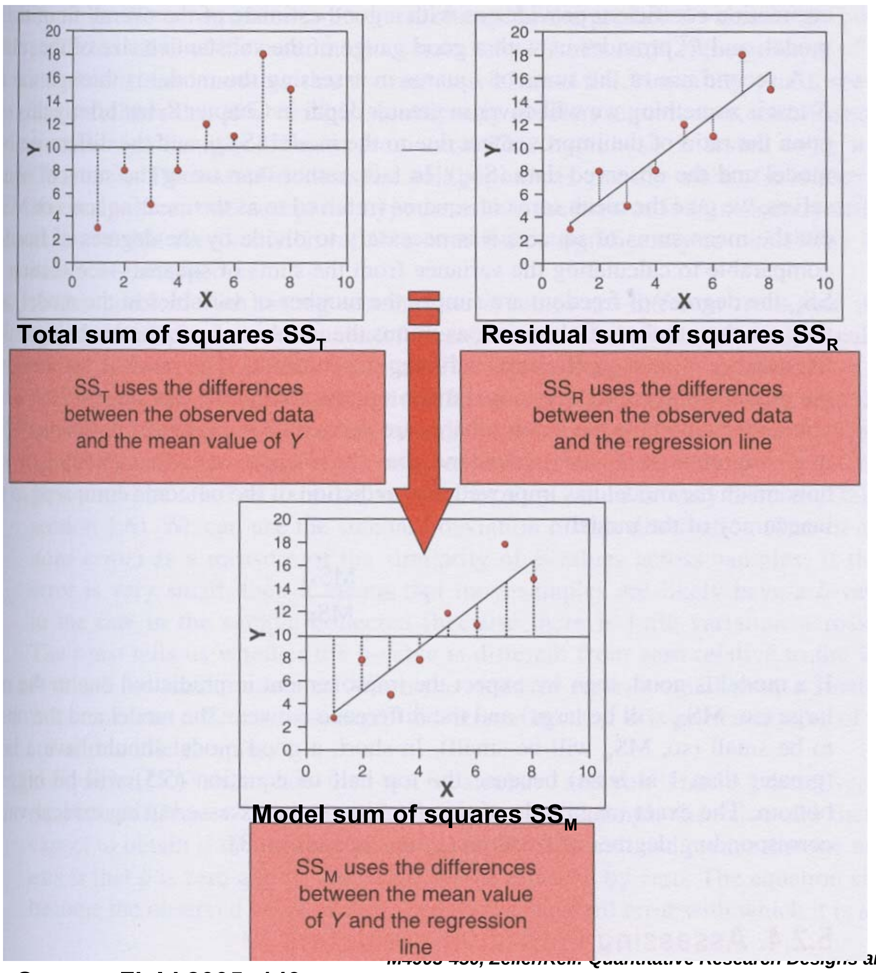
$$Y_i = ß_0 + ßX_i + e_i$$

Y_i = Dependent variable for cases i = 1 to n where n = total number of cases used in regression model

$ß_0$ = Intercept on y-axis = constant term

ß = Regression coefficient (dY/dX) = change in Y due to a one-unit change in X = slope of regression line

e_i = Residual = unexplained difference between the observed value of Y_i and the predicted value at point X_i



Source: Field 2005: 149

Simple regression (2)

➔ A regression function determines the slope dY/dX in such a way that the sum of the squared distances between the regression line and the actual values is minimized (*Ordinary Least Squares* [OLS] method).

Assessing the *goodness-of-fit* using the *Coefficient of Determination* (R^2):

$$R^2 = SS_M/SS_T$$

= Share of variation in the outcome variable that can be explained by the model

Simple regression (3)

F-test to assess overall model applicability:

➔ **The F-test is based on the ratio of the improvement in explaining the variance in Y due to the model (SS_M) and the difference between the model and the observed data (SS_R).**

➔ **However, it doesn't directly use the sum of squares (SS), but the *mean sum of squares* (MS), which is the SS divided by the degrees of freedom (df):**

- **$MS_M = SS_M$/ # predictor variables in the model**
- **$MS_R = SS_R$/ (N - # parameters estimated by the model, i.e., the coefficients of the predictor variables and the constant)**

➔ **$F = MS_M/MS_R$**

➔ **Measure of how much the model has improved the prediction of the outcome variable compared to the level of inaccuracy in the model**

➔ **A good model should have a large F-ratio!**

➔ **The F-test compares the model F-ratio to critical values for the corresponding degrees of freedom (tabulated in all statistics books).**

Simple regression (4)

T-test on the regression coefficient dY/dX:

➔ **Since our estimated regression coefficient is usually based on observations from a random sample rather than the whole population of interest, a different sample out of that population would most likely have led to a different value of the regression coefficient.**

➔ **We have to test the null-hypothesis that the 'true' regression coefficient is actually zero, i.e., that there is no linear relationship between X and Y, and our estimate deviates from zero by chance alone.**

➔ **All standard statistics programs calculate the t-value for every regression coefficient. If the t-value exceeds a certain critical value (given a pre-specified significance level), the null-hypothesis is rejected.**

Simple regression (5)

Deriving a point elasticity from the regression coefficient:

- ➔ **From the regression coefficient, we can calculate a *point elasticity (E)*, i.e., an unitless measure of the relative strength of the effect of the regressor on the dependent variable.**
- ➔ **Formula: E = (dY/dX)/(Y/X) = (dY/Y)/(dX/X), where X and Y are the sample means.**
- ➔ **Interpretation: Starting from the mean value, a 1% increase in the value of X will increase/decrease Y by E%.**
- ➔ **Example: Price elasticity of demand = percentage change in quantity demanded (Y) that occurs in response to a percentage change in price (X). If, in response to a 10% fall in the price of a good, the quantity demanded increased by 5%, the price elasticity of demand would be 5%/(− 10%) = −0.5**

SPSS Output for simple regression (1)

$Y_i = ß_0 + ßX_i + e_i$

Y = per person expenditure on clothes and footwear

X = Per adult value of land holdings

ß = dY/dX = regression coefficient where dX = 1.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,234[a]	,055	,053	2547,17590

a. Predictors: (Constant), Per adult value of land holdings

- ▪ **R-Squared = 0.055**
- ➔ **5.5 % of variance in Y is explained by X.**
- ▪ **Adjusted R-Squared = 0.053**
- ➔ **Very close to R-Squared**
- ➔ **The model does not contain redundant variables.**

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2E+008	1	187488340,0	28,897	,000[a]
	Residual	3E+009	498	6488105,054		
	Total	3E+009	499			

a. Predictors: (Constant), Per adult value of land holdings

b. Dependent Variable: Per person expenditure on clothes and footwear (Rupiah)

- ➔ **The F-test is highly significant**
- ➔ **Overall, our model constitutes a significant improvement in predicting Y as compared to using the mean of Y as the simplest possible statistical model.**

SPSS Output for simple regression (2)

➔ **The regression coefficient in the example output below is 0.002, which means that if the land value owned by an adult in the household increases by 1 monetary unit, the clothing expenditure increases by 0.002 units. In the SPSS-Output, the regression coefficient is denoted as B (N = 500).**

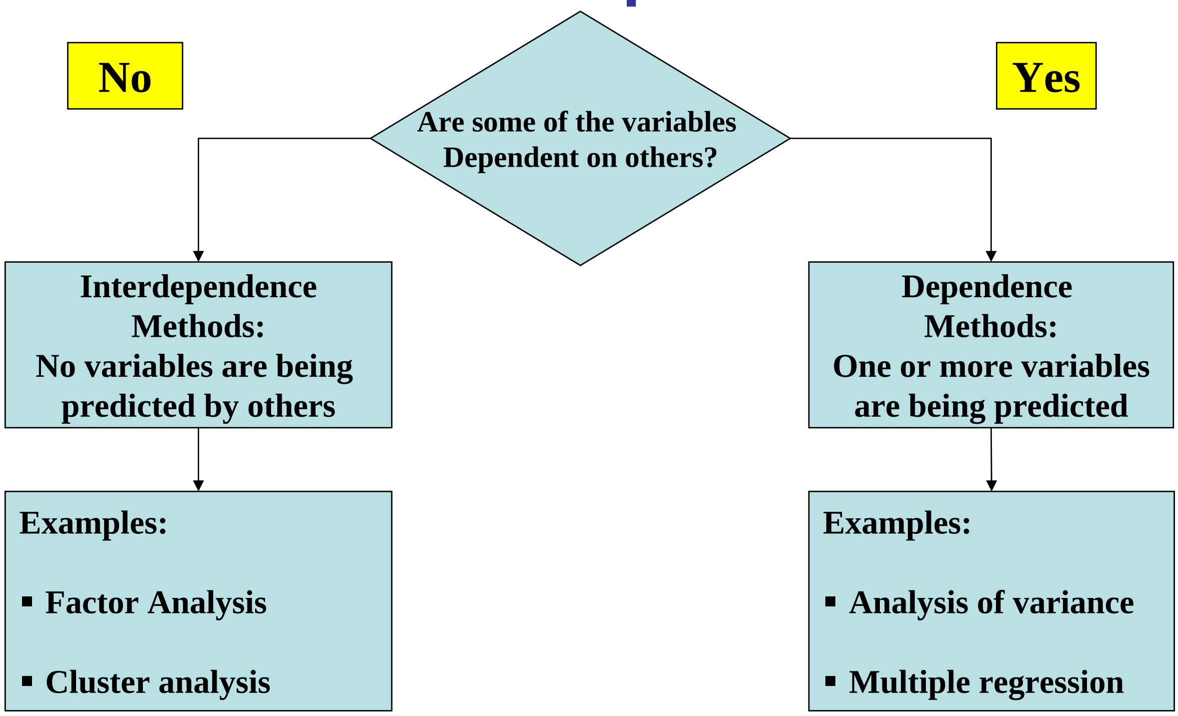
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2367,753	133,388		17,751	,000
	Per adult value of land holdings	,002	,000	,234	5,376	,000

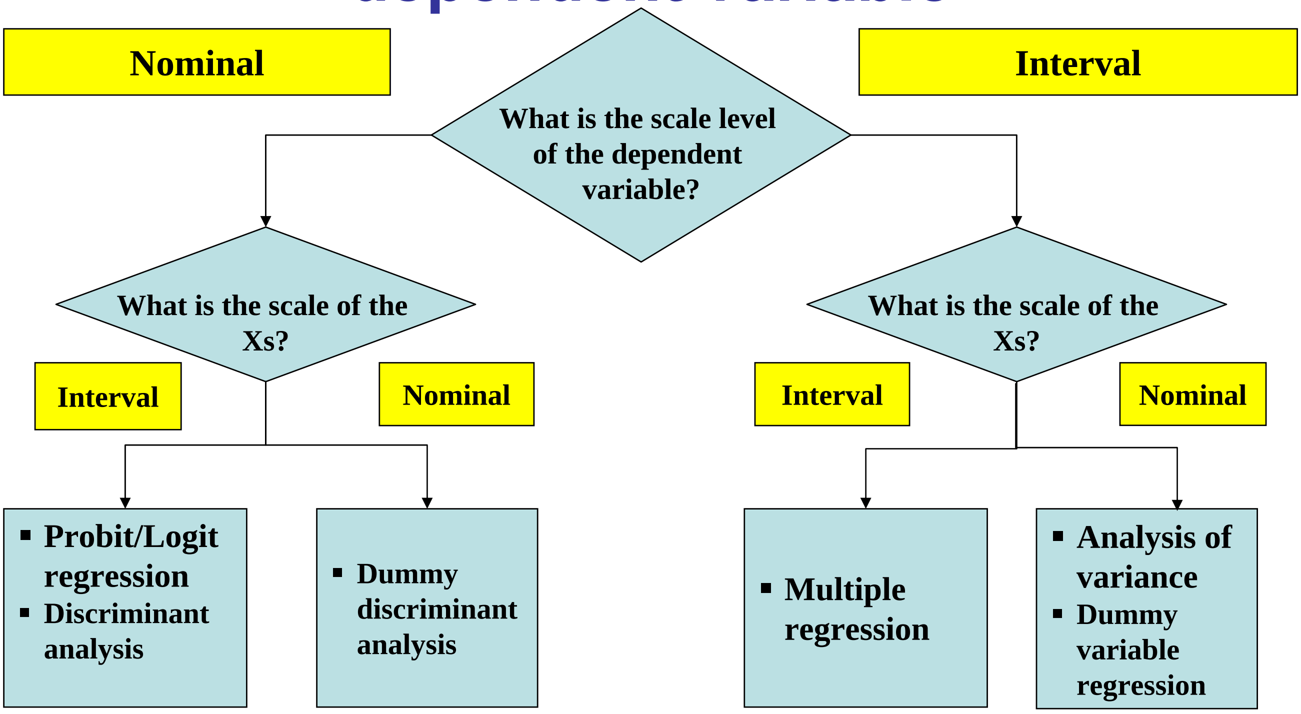
a. Dependent Variable: Per person expenditure on clothes and footwear (Rupiah)

Test value t = 5.376 > critical value = 1.96
(from t-distribution table for *P* = 0.05, df = 498 = N - # predictor variables - 1)
➔ Reject the null-hypothesis (i.e., B = 0) at *P* < 0.001!

Overview of multivariate analysis techniques



Dependent methods with one dependent variable



Analysis of variance (ANOVA)

- We use this to test the null-hypothesis that **more than two sample means** are equal. This technique examines the variability of the observations within each group (for example different ethnic groups) as well as the variability between the group means.
- The simplest case is the so-called ***one-way ANOVA*** procedure. In this case, only one variable is used to classify into the different groups. An F-value is calculated. If this value exceeds some critical level, the null-hypothesis is rejected.
- However, a significant F-test only tells us that there is a statistically significant difference between **any** of the sample means. To find out which means differ from each other, we have to apply so-called ***post-hoc tests*** for pairwise comparisons. The choice of post-hoc test depends on whether or not the variances in the groups that are being compared are equal (this is tested by ***Levene's test***).
- SPSS allows us to run Levene's test together with the one-way ANOVA and then select among various post-hoc tests that differ in terms of their control over ***Type I and Type II errors***.
- If we have more than one classifying variable, we need to use ***Factorial ANOVA*** (e.g., means of several ethnic groups, further differentiated by marital status).

SPSS Output for a one-way ANOVA with post-hoc test (1)

Example: Do clothing expenditures differ between wealth categories, defined as 'poorest', 'less poor' and 'least poor'? (This categorization is based on terciles of a poverty index)

➔ The null-hypothesis is that the mean expenditures in the three groups are not different.

Descriptives

Per person expenditure on clothes and footwear

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
poorest	166	1288,0984	966,20755	74,99225	1140,0303	1436,1666	,00	5400,00
less poor	167	2429,6868	1536,92666	118,93096	2194,8745	2664,4990	156,25	12500,00
least poor	166	4520,0355	3460,08154	268,55441	3989,7894	5050,2816	700,00	26666,67
Total	499	2745,3065	2618,10707	117,20257	2515,0340	2975,5789	,00	26666,67

ANOVA

Per person expenditure on clothes and footwear

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	9E+008	2	445986744,9	87,727	,000
Within Groups	3E+009	496	5083790,018		
Total	3E+009	498			

➔ The significant F-test indicates that there is a difference between the three mean values!

SPSS Output for a one-way ANOVA with post-hoc test (2)

Test of Homogeneity of Variances

Per person expenditure on clothes and footwear

Levene Statistic	df1	df2	Sig.
44,752	2	496	,000

➔Levene's test rejects the null-hypothesis of homogeneity of variances between the three groups at *P* < 0.001!
➔We select a post-hoc test for pairwise comparisons that is suitable for unequal variances between groups
➔e.g., the Games-Howell test

Multiple Comparisons

Dependent Variable: Per person expenditure on clothes and footwear

Games-Howell

(I) Poverty terciles	(J) Poverty terciles	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
poorest	less poor	-1141,5883*	140,60018	,000	-1472,8785	-810,2982
	least poor	-3231,9371*	278,82846	,000	-3890,5776	-2573,2965
less poor	poorest	1141,58835*	140,60018	,000	810,2982	1472,8785
	least poor	-2090,3487*	293,71082	,000	-2783,2612	-1397,4363
least poor	poorest	3231,93707*	278,82846	,000	2573,2965	3890,5776
	less poor	2090,34872*	293,71082	,000	1397,4363	2783,2612

*. The mean difference is significant at the .05 level.

➔We reject the null-hypothesis of equality of means between all pairs at *P* < 0.001!

Structure of a regression model (1)

Generic model with two independent variables:

$\mathbf{Y_i = ß_0 + ß_1 X_{1i} + ß_2 X_{2i} + e_i}$

Y_i = Dependent variable for case i =1 to n where n = total number of cases used in regression model

$ß_0$ = Intercept on y-axis

$ß_1$ = Regression coefficient (dY/dX_1) or slope = Change in Y with a change of X_1 by one unit. (Listed as B in SPSS output)

$ß_2$ = Regression coefficient (dY/dX_2) or slope = Change in Y with a change of X_2 by one unit.

e_i = Residual (unexplained difference between the observed value of Y_i and the predicted value at the point X_i)

Structure of a regression model (2)

To be able to draw conclusions about the population based on an OLS multiple regression analysis performed on sample data, **the following assumptions must be met**:

1) **Variable types**: All predictor variables must be interval variables or dummy variables (binary nominal variables that can take on the value of either 0 or 1). The outcome variable must be a continuous and unbounded interval variable.
2) **The average value of Y is given by a linear combination of the predictor variables** (plus a constant term), i.e., **on the average, the error term is zero** (i.e., in the above model $E(Y) = ß_0 + ß_1X_1 + ß_2X_2$).
3) **Error terms should be normally distributed** (with a mean of zero).

Structure of a regression model (3)

4) **No perfect multicollinearity**: Predictor variables must not correlate with each other too highly (e.g., |r| > 0.8).

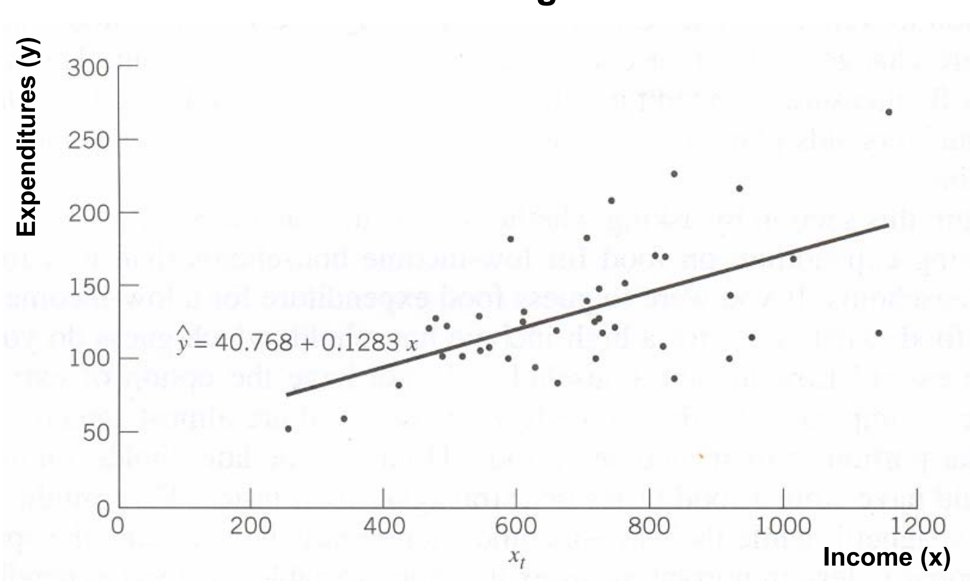
5) **Homoskedasticity**: At each level of the predictor variable(s), the variance of the residuals (error terms) should be constant.

6) **No autocorrelation**: For any two observations, the residual terms should be uncorrelated (independent errors).

➔ There are a number of **tests** that need to be performed in order to test whether the model specification is appropriate, i.e., the above assumptions are met. A good introductory description is provided in the SPSS base manual, including practical guidelines for tests. For further introductory literature, see for example Field (2005) or Kennedy (2003).

Heteroskedasticity (1)

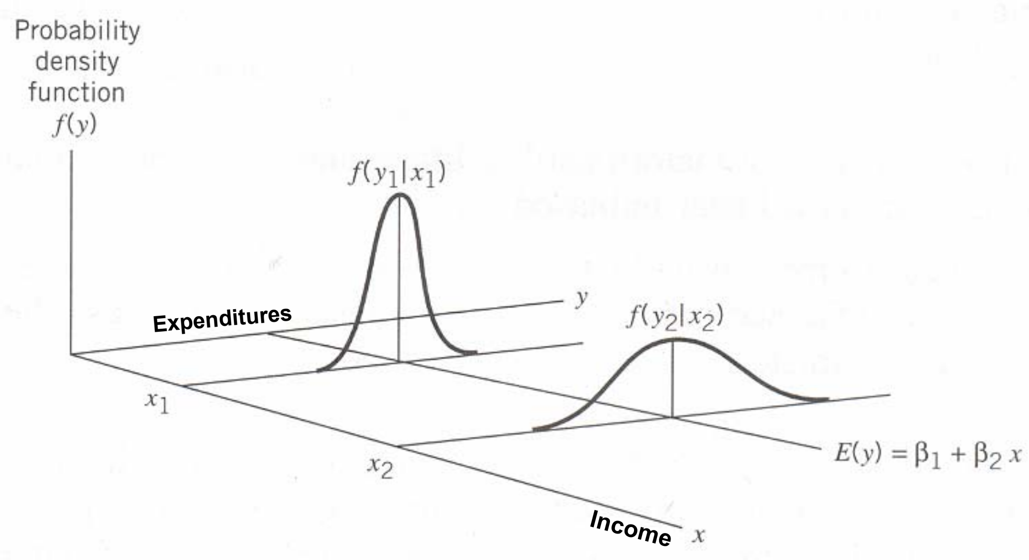
Example: Effect of income on expenditures: with increasing income, observed expenditure levels tend to deviate more and more from an estimated regression function:



Least squares estimated expenditure function and observed data points
Source: Adapted from Hill et al. 2001: 236

Heteroskedasticity (2)

➔ **This means that the error variances are not constant across the range of the predictor variable:**



Heteroskedastic errors
Source: Adapted from Hill et al. 2001: 237

Heteroskedasticity (3)

Consequences of heteroskedastic errors:

- The least squares (LS) estimator is still linear and unbiased, but not the 'best' among linear and unbiased estimators (i.e., it is only 'LUE' but not 'BLUE').
- The standard errors usually computed for the LS estimators are incorrect; in particular, they may be too small, thus overstating the precision of the estimates.

➔ Confidence intervals and, consequently, hypothesis tests may be misleading; in particular, t-tests on regression coefficients may erroneously reject the null-hypothesis.

Heteroskedasticity (4)

Detecting heteroskedasticity:

- Visual inspection by plotting the residuals against the fitted values of the regression model.
- ➔ If the errors are homoskedastic the graph should not exhibit any systematic pattern.
- Use of formal tests, e.g., Goldfeld-Quandt test, Breusch-Pagan test (see the literature for details on these tests).

Remedial measures:

- Use of White's robust standard errors.
- Use of the Generalized Least Squares (GLS) estimator.
- ➔ Can be viewed as a weighted LS estimator, which takes the value of the predictor variable X into account.
- ➔ E.g., if large values of X are associated with larger error variances than small values, they would be weighted more lightly (consult the literature for details on GLS).

SPSS output of a multiple regression (1)

Model structure:

$Y_i = ß_0 + ß_1X_{1i} + ß_2X_{2i} + ß_3X_{3i} + e_i$

Y = per person expenditure on clothes and footwear

X_1 = Total number of adults and children in the household

X_2 = Per adult value of land holdings

X_3 = Average age of adults in the household

SPSS output:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,349[a]	,122	,116	2460,58193

a. Predictors: (Constant), Average age of adults in household, Per adult value of land holdings, Total number of adults and children in household

- R-Squared = 0.122
- ➔ 12.2% of variance in Y is explained by X.
- Adjusted R-Squared = 0.116
- ➔ Close to R-Squared
- ➔ The model does not contain redundant variables.

Note: Although all variables included in the model have statistically highly significant regression coefficients (see overleaf), the total share of variance explained is very low!

SPSS output of a multiple regression (2)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4E+008	3	138516930,1	22,878	,000[a]
	Residual	3E+009	496	6054463,441		
	Total	3E+009	499			

a. Predictors: (Constant), Average age of adults in household, Per adult value of land holdings, Total number of adults and children in household

b. Dependent Variable: Per person expenditure on clothes and footwear (Rupiah)

- ➔ **The F-test is highly significant**
- ➔ **Overall, our model constitutes a significant improvement in predicting Y as compared to using the mean of Y as the simplest possible statistical model.**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5067,147	560,035		9,048	,000
	Total number of adults and children in household	-316,032	53,626	-,251	-5,893	,000
	Per adult value of land holdings	,003	,000	,259	6,116	,000
	Average age of adults in household	-33,628	13,369	-,107	-2,515	,012

a. Dependent Variable: Per person expenditure on clothes and footwear (Rupiah)

➔ **All t-tests on the regression coefficients reject the null-hypothesis (i.e., $B_i = 0$) at $P < 0.05$ (or lower)!**

Dummy variables (1)

- **The regression model $Y_i = ß_0 + ß_1X_{1i} + ß_2X_{2i} + ß_3X_{3i} + e_i$ assumes that the model structure and the estimated parameters $ß_k$ are identical for each observation i.**
- **However, we often encounter situations in which this assumption does not hold, due to qualitative differences between observations.**
- **Dummy variables are instruments for capturing such qualitative differences, e.g. with regard to gender, ethnicity, geographic region etc.**
- **Dummy variables only take on two values, '1' if a certain characteristic is present, and '0' otherwise.**
- **The LS estimator's properties are not affected by the fact that a dummy variable consists only of zeros and ones; a dummy variable is treated as any other explanatory variable.**

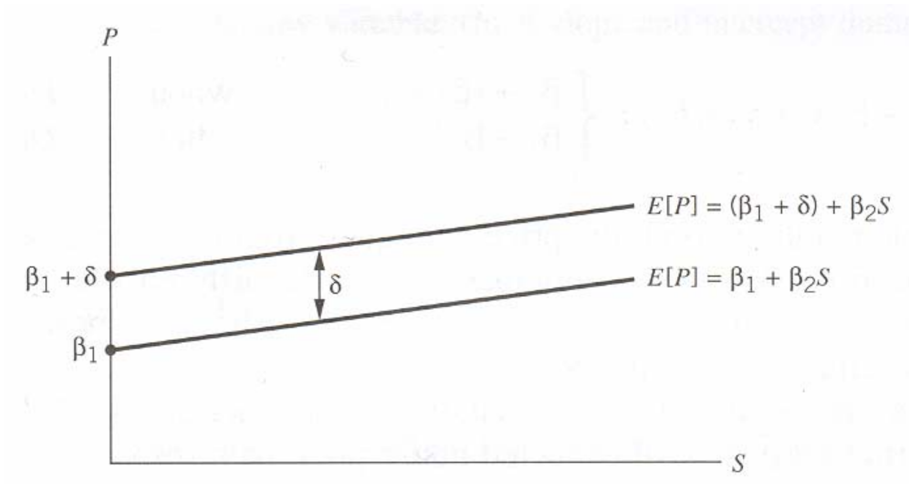
Dummy variables (2)

Example (taken from Hill et al. 2001):

- **Let's consider the simple regression model**
  $P_i = ß_1 + ß_2S_i + e_i$
  where P is the price and S the size of a house.
- **$ß_1$ can be interpreted as the value of the plot of land alone (standardized in size), and $ß_2$ as the value of an additional square metre of living area.**
- **However, in real estate prices, location matters, too. Let's therefore define the dummy variable D_i, which takes on the value of 1 if the property is in a particularly desirable neighborhood, and the value of 0 otherwise. We thus extend the model to**
  $P_i = ß_1 + \delta D_i + ß_2S_i + e_i$

Intercept dummy variables

- **Adding D_i to the model creates a parallel shift in the relationship by the amount δ, the 'location premium'.**
- **This kind of dummy variable that captures a shift in the intercept due to some qualitative factor is called an intercept dummy variable.**



An intercept dummy variable
Source: Hill et al. 2001: 201

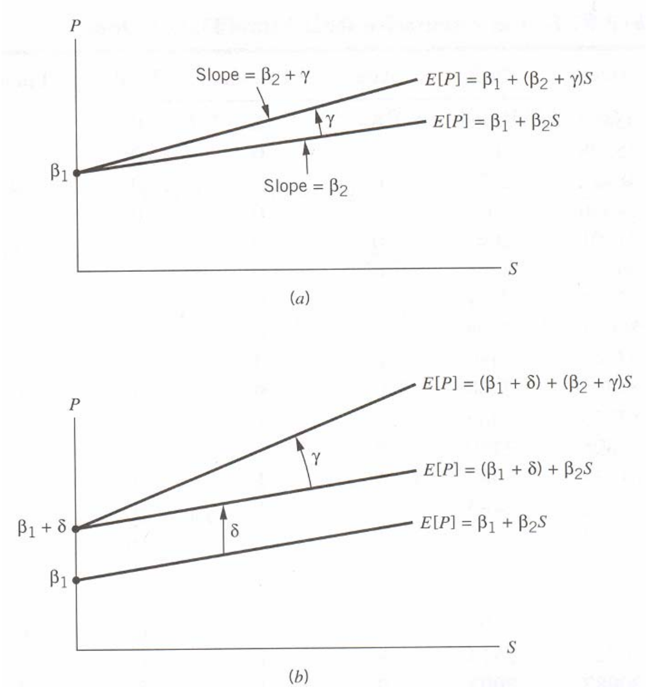
Slope dummy variables

Instead of assuming that location causes a constant 'premium' resulting in a shift of the intercept, we could specify an alternative model in which the price per square metre (i.e. the slope of the regression line) is increased by adding the product of house size and the dummy variable (= interaction variable):

$P_i = ß_1 + ß_2S_i + γ(S_iD_i) + e_i$

Finally, both possible effects could be accommodated in a single model:

$P_i = ß_1 + δD_i + ß_2S_i + γ(S_iD_i) + e_i$



(a) A slope dummy variable (b) A slope and intercept dummy variable Source: Hill et al. 2001: 203

Binary choice models

- In economics, the outcome variables of interest are not always continuous in nature; often, individuals or firms make *choices of an 'either-or' nature*.
- Examples: Adoption/non-adoption of a technology, participation/non-participation in a program.
- In econometrics, a binary variable is used to capture choices of this kind, i.e., the variable takes on the value of '1' if a certain outcome is chosen, and '0' otherwise.

➔ Use of a *dummy variable as the dependent variable*, rather than an explanatory variable.

➔ This has important implications on the kind of statistical model to apply.

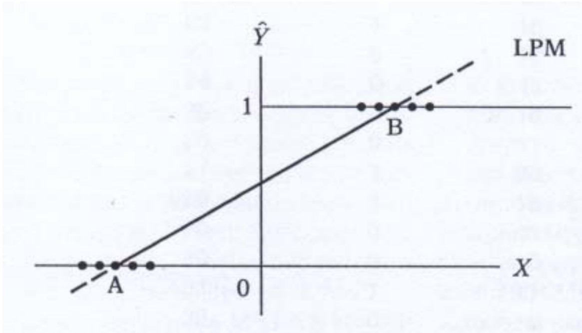
The Linear Probability Model

- Let's define a binary variable y that takes on the value of 1 if a farmer has adopted a certain technology, and 0 otherwise.
- The probability function of this variable is $f(y) = p^y (1-p)^{1-y}$, where y = 0, 1 and p is the probability that y = 1.
- The simple Linear Probability Model (LPM) explaining the choice variable y is given by $y_i = ß_1 + ß_2x_i + e_i = E(y_i) + e_i = p + e_i$, where x is a suitable explanatory variable.

➔ By estimating this linear model using OLS, we assume that increases in x have a constant effect on the probability that a farmer will adopt: $dp/dx = ß_2$

➔ When using this model to predict behavior by substituting alternative values of x, we can easily obtain predicted probabilities of adoption < 0 or > 1, which do not make sense.

➔ We need to employ a *nonlinear model* to keep the choice probability strictly within the interval [0, 1].



The Linear Probability Model (LPM)
Source: Adapted from Gujarati 2003: 587

The Probit Model

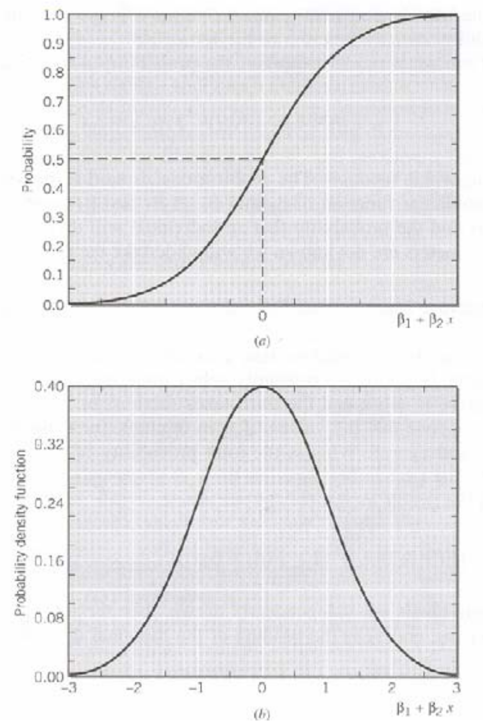
- To keep the choice probability p strictly within the interval [0, 1], a nonlinear, S-shaped relationship between x and p can be employed.
- In the Probit Model, this is the standard normal probability distribution.
- The slope of the cumulative distribution function (CDF) represents the change in probability, given a unit change in x, which is not constant as in the LPM.
- The probit function is given by

$$F(z) = P[Z \le z] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-0.5z^2} dz$$

- The Probit Model expresses the probability p that y takes on the value of 1 to be

$$p = P[Z \le \beta_1 + \beta_2 x] = F(\beta_1 + \beta_2 x)$$

where F is the probit function.



(a) Standard normal cumulative distribution function (CDF)
(b) Standard normal probability density function (PDF)
Source: Hill et al. 2001: 371

Estimation of the Probit Model

- Since the Probit Model is a nonlinear function of $ß_1$ and $ß_2$, we cannot use the LS procedure for their estimation, but employ the *maximum likelihood* (ML) method.
- The ML procedure chooses values for $ß_1$ and $ß_2$ in such a way that the likelihood of observing the sample data is maximized (i.e., it maximizes the *likelihood function*).
- Because of the complexity of the probit function F, the maximization is done by computers using an *iteration process*.
- It can be shown that *for large samples* the ML estimator is normally distributed, consistent, and 'best', i.e., it has the smallest variance.

Interpretation of the Probit Model

- The probit estimates (regression coefficients) are not directly interpretable as the change in probability p that y = 1 for a one-unit change in each independent variable x!
- To calculate this *marginal effect*, we have to consider the derivative of the Probit Model, which yields (using the chain rule)

$$\frac{dp}{dx} = \frac{dF(t)}{dt} \cdot \frac{dt}{dx} = f(\beta_1 + \beta_2 x)\beta_2$$

where $t = \beta_1 + \beta_2 x$ and $f(\beta_1 + \beta_2 x)$ is the standard normal PDF evaluated at $\beta_1 + \beta_2 x$

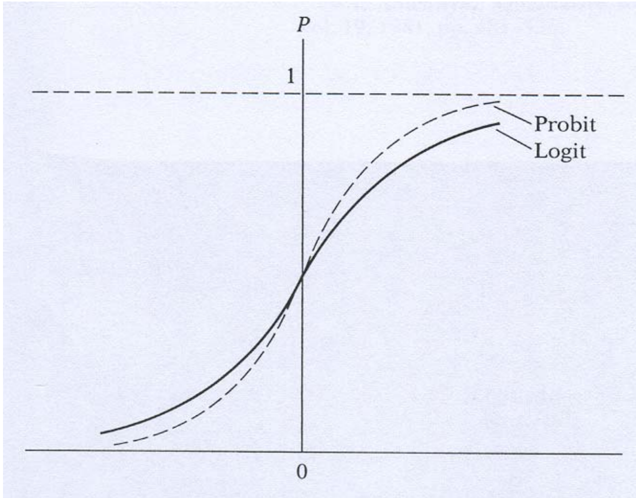
➔ It thus becomes apparent that the marginal effect will depend on the particular values of *all independent variables*!

➔ All software packages that include probit estimation can produce the marginal effects of a one-unit change in each independent interval-scale variable (starting from its mean!), and the discrete change from 0 to 1 for dummy variables, taking all other independent variables at their means.

➔ Care has to be taken to interpret the probit output in exactly this way!

Probit and Logit Models

- An alternative to the Probit Model is the Logit Model.
- The two models differ only in the particular, underlying S-shaped function to constrain probabilities to the [0, 1] interval.
- Instead of the standard normal distribution, the Logit Model is based on the logistic distribution, which has slightly 'fatter' tails, i.e., the conditional probability p_i approaches 0 and 1 at a slower rate.
- Both models usually produce very similar results; there is no compelling reason to choose one over the other.



Probit and Logit cumulative distributions
Source: Gujarati 2003: 614

Sample Selection Bias (1)

- **Sample selection bias** can occur because of (1) decisions taken by the researcher, or (2) decisions taken by the observed individuals, both leading to data that are not representative of the alleged population.
- **Example for case (1):** In 1945, opinion polls dramatically failed to predict the presidential win of Harry S. Truman. Why? Polls were conducted via telephone, which at that time were more likely to be owned by relatively wealthy people; at the same time, wealthy people were more likely to vote for the opponent, Dewey.

➔ The unmeasured variable 'wealth' affected both the survey answer, and the probability of being in the sample, thus creating a misleading result.

Sample Selection Bias (2)

- **Example 1 for case (2):** Suppose you have conducted a mail survey examining the determinants of the time spent watching TV.

➔ The unmeasured variable 'laziness' may affect both the time spent watching TV, and the probability of returning the questionnaire (and thus being in the sample).

- **Example 2 for case (2):** Sample selection bias may be much more subtle than in the previous examples: For instance, observations on hours worked are available only for people whose wage exceeds their reservation wage.

➔ The problem is that we often wish to draw conclusions about a wider population, not just the sub-population from which the data are derived. But our sample is not representative, it is afflicted by *self selection bias*.

Sample Selection Bias (3)

- **In general:** As stated in the section on research designs, true experiments are almost impossible to conduct in the social sciences. In particular, a random assignment of human beings to a 'treatment group' (in the broadest sense) and a 'control group' is problematic for ethical and practical reasons.
- Instead, we usually face the situation that our observation units have *selected themselves* into the treatment and control groups (Ex-post facto research design).
- If we are interested in measuring variables that are only observable within the treatment group, our data are not (necessarily) representative of the population we drew the original random sample from.

Sample Selection Bias (4)

➔ If unobserved or unobservable factors influence both the probability that an observation unit appears in our sample and the variables of interest, our estimates will be afflicted by *sample selection bias*.

➔ In such cases, the estimation procedure applied must take the sample selection phenomenon into account!

The Heckman Procedure (1)

- The two-step Heckman procedure is a popular tool for correcting sample selection bias (Heckman 1979).
- This approach involves (1) a *selection equation* identifying determinants of whether of not an observation unit appears in the sample:

$$z^* = \gamma w + \nu$$

- And (2) a *regression equation of primary interest* based on the non-random sub-sample:

$y = \beta x + \varepsilon$, where

w, x = vectors of exogenous regressors

γ , β = vectors of parameters

ν , ε = error terms

The Heckman Procedure (2)

➔ **1. Step:** Run a Probit Model to identify factors determining whether of not an observation unit appears in the sample.

$$z^* = \gamma w + \nu$$

➔ In situations of self-selection z* can be interpreted as the difference in expected returns between two alternative outcomes, e.g., adoption and non-adoption of an improved wheat variety.

➔ Since z* is not directly observable, a binary outcome variable z is defined, that can either take on the value of 0 or 1:

$z = \gamma w + \nu$, where

$z = 1$ if $z^* > 0$, and $z = 0$ otherwise.

The Heckman Procedure (3)

➔ **The probability that an observation appears in the sample can be expressed as:**

$\text{Prob}(z = 1) = \text{Prob}(z^* > 0) = \text{Prob}(\nu > -\gamma w) = \text{Prob}(\gamma w) = \Phi(\gamma w)$,

where Φ is the cumulative distribution at γw.

➔ **In the regression equation of primary interest ($y = \beta x + \varepsilon$), y is only observed if $z = 1$. Furthermore, the error terms of the two equations are assumed to follow a bivariate Normal distribution with mean 0, standard deviation σ, and correlation ρ. Then:**

$E[y|z=1] = \beta x + \rho\sigma\lambda(\gamma w)$, where

$\lambda = \phi(\gamma w)/\Phi(\gamma w)$, where

ϕ is the density function at γw.

➔ λ **is the *Inverse Mills Ratio* (IMR), and** $\rho\sigma$ **is the regression coefficient on the IMR,** β_λ

The Heckman Procedure (4)

➔ **2. Step:** Run an OLS regression on the dependent variable of interest, including the IMR as an additional explanatory variable:

$$y = \beta x \ + \beta_\lambda \lambda + \varepsilon$$

➔ The IMR can be interpreted as a variable that captures all unobserved and unobservable characteristics affecting both the probability of being in the observed sub-sample and the outcome variable of interest.

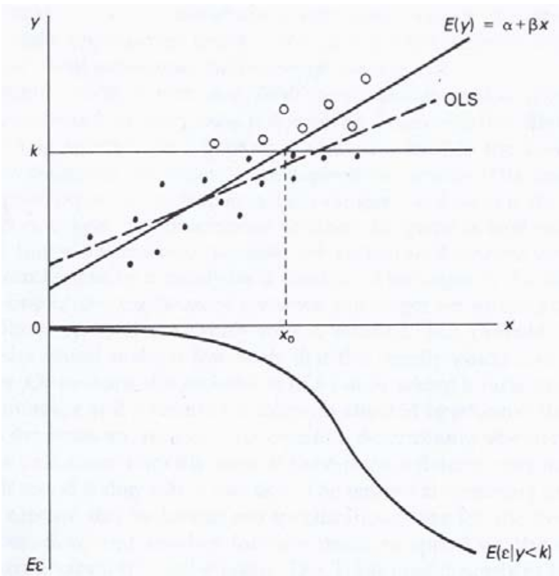
➔ By including the IMR in the second-stage equation, the sample selection bias is corrected for, and OLS can safely be used for the estimation of β.

Graphical illustration of the IMR

- **Assume that values of y in the upper half of the graph are only observable if y < k (e.g., y may be the demand for tickets to soccer games, and k may be the capacity of the stadium; hence, in cases when y ≥ k, we observe k).**
- **If we regress y on x using OLS (i.e., assuming E(ε) = 0), the slope of the resulting regression line (dashed) will be too flat.**
- **The lower half of the graph shows the expected error term [E(ε)] for a given value of x, under the condition that y is being observed: for instance, to observe the corresponding y-value of x_0, ε has to be 0 or negative since, otherwise, y would exceed k and would thus be unobserved.**

➔ **The error term is correlated with the explanatory variable, causing the bias in the OLS estimates.**

➔ **The first stage of the Heckman procedure estimates the expected value of the error (the IMR), and the second stage runs the regression of interest including the IMR as an additional explanatory variable, thus correcting the bias.**



The rationale behind the Heckman procedure

Source: Adapted from Kennedy 2003: 282

Note: The problem illustrated in the figure can be better resolved using a so-called Tobit model (not within the scope of this module) using maximum likelihood estimation (Tobin 1958). It is therefore used for illustrative purposes only.

Further reading for interested students

- Heckman, J.J. (1979) Sample selection bias as a specification error. *Econometrica* 47, 153-161.
- Tobin, J. (1958) Estimation of relationships for limited dependent variables. *Econometrica* 26, 24-36.