# Quantitative Research Methods

Stefan Schwarze

University of Goettingen
Department of Agricultural Economics and Rural Development
Goettingen, Germany

# Why do we need economic data for conservation planning?

- Vast expansion of human activity during the last century

  => widespread conversion of natural habitat
  => many species are at risk of extinction
- Large-scale anthropogenic threats to biodiversity, but only limited resources devoted to conservation.
- Where to use the limited resources for conservation?

  => Maximize the conservation return on investment

Two domains for socio-economic assessments in conservation planning:

- Understand the threat to biodiverstiy

  => Why do people change their land use?
- Assess the costs of conservation

  => Which costs occur?

# Costs of conservation

Acquisition and transaction costs

- Costs involved buying the land

Management costs

- Costs for managing the park (staff, vehicles, offices, etc.)

Damage costs

- Damages of crops and livestock due to wild animals

Opportunity costs

- Potential economic benefits from using the park area for economic activities instead of protecting it

# Subject Matter

Focus is on quantitative methods of socio-economic analysis

- Design of socio-economic research
- Methods for data collection
- Methods of data analysis

## Aims

Ability to gather and analyze socio-economic land use data, particularly:

- appreciate the different methods of sampling commonly used
- apply best practices in questionnaire design
- calculate opportunity costs of conservation (gross margin analysis)

# Major References


Black, Thomas R. (1999):
Doing quantitative research in the social sciences. An Integrated approach to research design, measurement and statistics.
Sage Publications, London.

Ellis, Frank (2000):
Rural livelihoods and diversity in developing countries.
Oxford university Press, Oxford, GB.

Burns, Robert B. (2000):
Introduction to research methods.
Fourth edition. Sage Publications, London.

# Stages in the research process

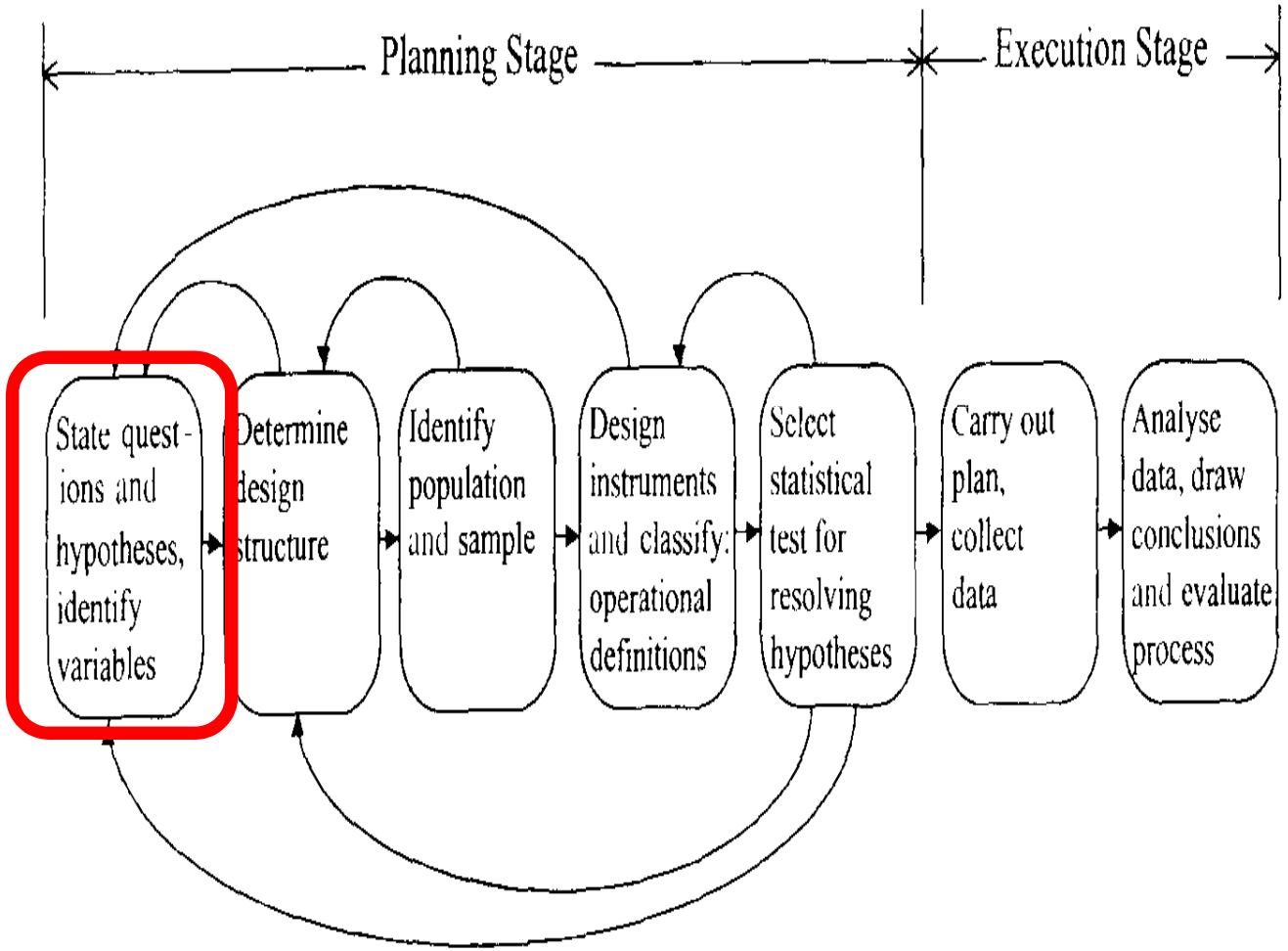



*FIGURE 2.1*
Stages of designing and carrying out a study, including iterations for modifications and improvements during planning

Source: Black, 1999

# Stage 1: Types of research questions

## 1. Descriptive: Question of the type ‘What is’?

*What is the percentage share of farmers growing grapes in the research area?*

No need for a theory, but theory helps you in guiding which variables you wish to describe.

For the remaining four types of research questions a theory/ conceptual framework is recommended, because we look on the relationship among variables.

# Stage 1: Types of research questions

**2. Explorative: Which characteristics relate to the observed outcomes?**

*Is there a relation between access to credit and farm incomes?*

**3. Explanatory: What are the causes of an observed outcome?**

*What is the impact of improved access to credit on farm incomes?*

**4. Predictive: What will happen if one variable changes?**

*How will the share of wheat in cultivated area change if the price of wheat increases by 10 %, all other factors held constant?*

# Stage 1: Research questions and hypotheses

**Research question => Hypothesis => Null hypothesis**

Statistics can tell us whether the outcomes we observe are due to some relationship OR simply by chance.

The null hypothesis is the hypothesis we want to nullify/refute

=> the outcomes did occur by chance

Statistical tests tell us with which error level we can reject the null hypothesis.

# Stage 1: Research questions and Hypothesis

**Example:**

**Research question:**

Is there a relationship between education level of a mother and the vaccination status of her children?

**Hypothesis $H_1$:**

There will be a positive relationship between mother's education and child vaccination.

**Null-Hypothesis $H_0$** for statistical test:

There will be no relation between mother's education and vaccination status.

# Theories and conceptual frameworks

The identification of relevant variables as well as the development of hypothesis should be based on a theoretical foundation and/or conceptual framework.

**What is a theory of land use (change)?**

It is a set of propositions used to understand

- the "what" of land use (change);
- the “where” of and use (change) and
- the "why" of this change.

**What is a conceptual framework?**

Based on theory and empirical evidence a conceptual framework visualises the relationship between your major variables.

# The Thünen rings

Von Thünen (1783 – 1850) developed a model of agricultural land use based on a profit maximizing farmer.

Profit (locational rent L) of a crop is given by: $L = Y(P - C) - YDF$

Y: Yield (in t/ha)

C: Costs of production (in €/t)

F: Transport cost (in €/t/km)

P: Market price of the crop (in €/t)

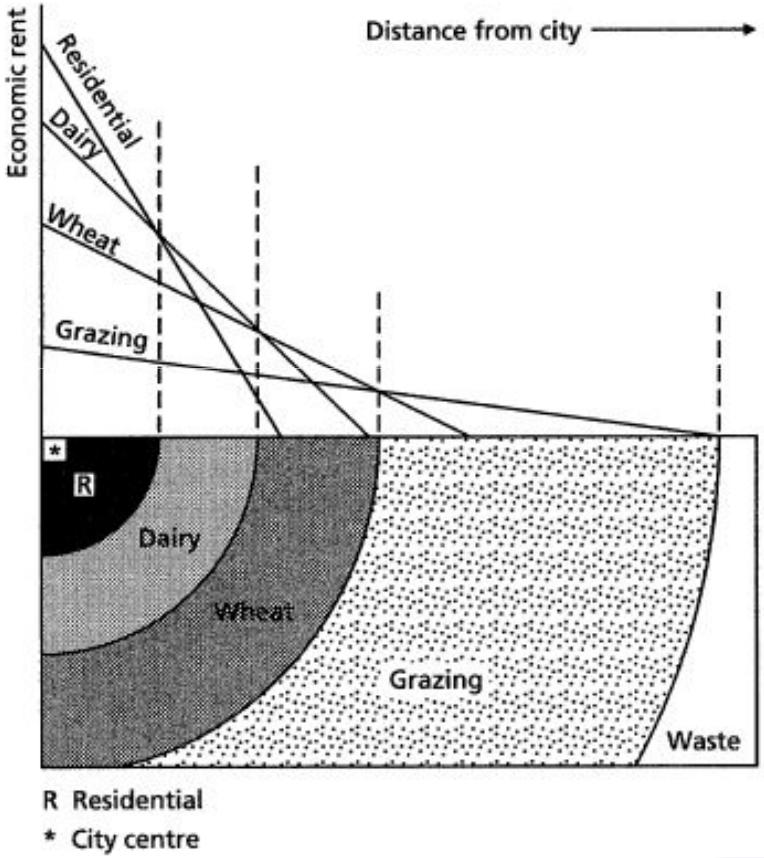
D: Distance from the market (in km)

Von Thünen concluded that the cultivation of a crop is only worthwhile within certain distances from the city. Beyond that, either:

- its profits drop to zero (because of transportation costs) or
- the profits earned by other crops are higher.

Since von Thünen referred transport costs directly to the market, circular land use zones arises - the Thünen rings.

# The Thünen rings




Source: www.answers.com

# A framework for livelihoods analysis

It illustrates the factors influencing the activity choice of rural households and, hence, also the livelihood outcomes.

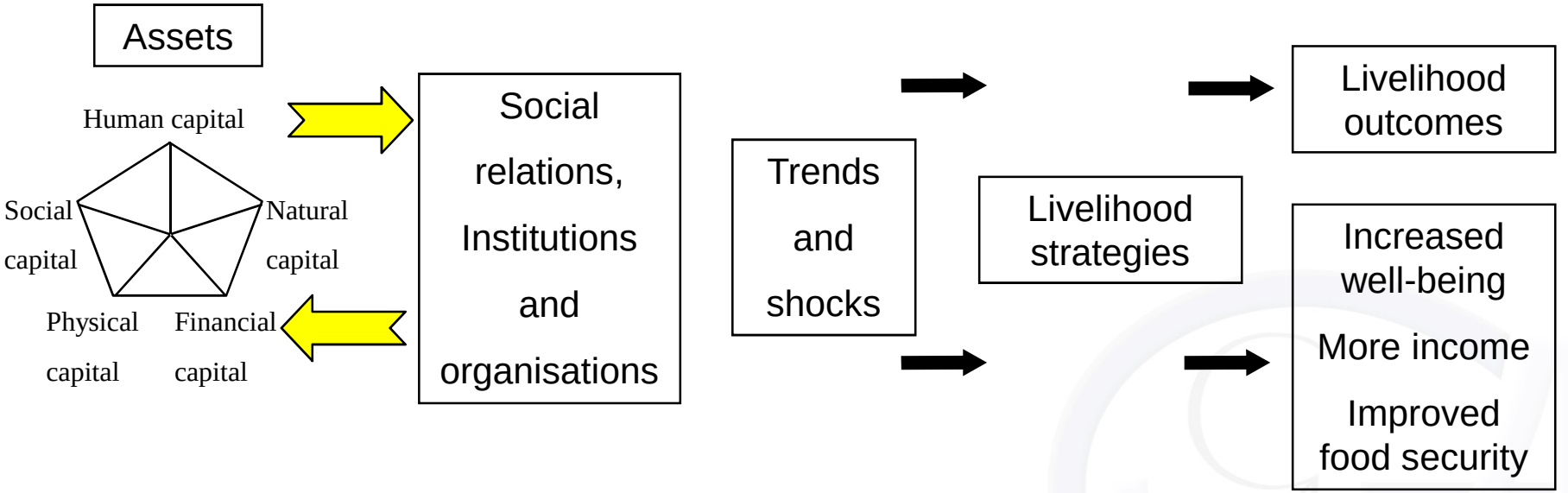
Particularly useful as a guide to micro policies (i.e. policies which directly affect rural households).

**What is a livelihood?**

"A livelihood comprises the assets, the activities, and the access to these (mediated by institutions and social relations) that together determine the living gained by the individual or household" (Ellis, 2000, p. 10).

# Livelihoods approach (1)

| Livelihood platform | Access modified by | In context of | Resulting in | In order to achieve |
|---|---|---|---|---|




Source: Ellis 2000

# Livelihoods approach (2)

**Assets**

- Resources owned, controlled, or in some other means accessed by the household.
- They are the building blocks to undertake production and engage in labor markets etc.

Asset categories:

Natural capital: land, water, and biological resources used by people

Physical capital: ‘man-made’ capital (tools, machinery etc.)

Human capital: labor available to the household

Financial capital: stock of money (savings/loans)

Social capital: social claims which households can draw by virtue of their belonging to a social group

# Livelihoods approach (3)

**Mediating processes**

- They translate a set of assets into a livelihood strategy.
- Factors that influence access to assets and their use.
- Distinction of factors which are endogenous to the social norms and structures and exogenous factors.

Process categories:

Social relations: gender, class, age ethnicity

Institutions: rules and customs, land tenure, markets

Organizations: NGOs, local administration, state agencies etc.

Trends: population, technological change, macro policies, world economic trends

Shocks: droughts, floods, pests, diseases, civil war

# Livelihoods approach (4)

**Livelihood strategies**

- Asset status mediated by factors results in livelihood strategies.
- Livelihood strategies are dynamic.
- These strategies are composed of activities that generate outcomes.

Resource based strategies: collection and cultivation of food/non-food, livestock keeping

Non-resource based strategies: trade and manufacturing, remittances and other transfers

Total household income refers to:

- both cash and in-kind contributions
- refers to net-income: gross income – cash expenses

# Stages in the research process

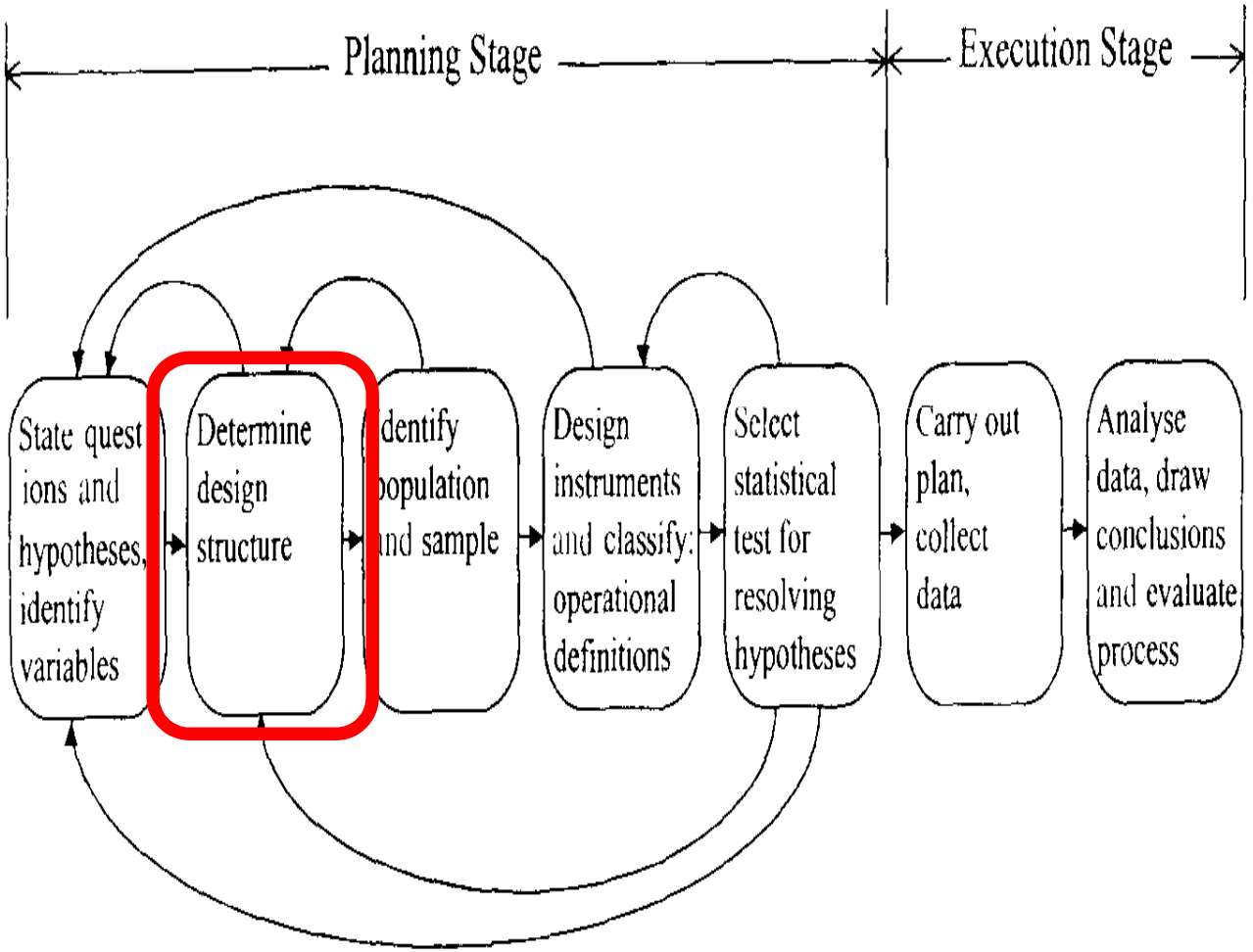



*FIGURE 2.1*
Stages of designing and carrying out a study, including iterations for modifications and improvements during planning

Source: Black, 1999

# Types of quantitative research designs (1)

**Correlation Studies**

- ask whether a relationship between two variables exist
- no pretence to establish causality

*Is the number of extension messages received related with adoption of improved wheat varieties?*

There is a relationship, but:

- Does extension influences adoption?
- Does adoption influence extension received?
- Does a third variable influence adoption and extension received?

# Types of quantitative research designs (2)

**Causal research**

- asks whether a certain variable influences another variable
- Tries to rule out any other influencing factors

*Is there a difference in adoption of improved wheat seed between farmers who received extension services and those who did not?*

=> a different research design is needed for causal research!

# Types of quantitative research designs (3)

A. Pre-experimental
- it is not suitable to test causality (purely descriptive), because no comparison is made to another group

B. Experimental
- the ideal design for testing causality
- but often not practicable in social science research

C. Quasi-experimental
- sometimes used, and better than D in establishing causality

D. Ex-post facto
- most often used for causal socio-economic research

# Experimental research design

General characteristics

1. True experiments always have two groups:
   - one group is subjected to the treatment/policy,
   - the other not (control group).
2. Random sampling and random assignment to the two groups.

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

Problems using it in social sciences:

1. Ethical, time or budget reasons make random assignment difficult
   => all farmers assigned to the treatment group MUST go to the extension meeting, even if they do not wish to.
   => all other farmers are excluded from the service, even if they want to participate.

# Experimental research design (2)

Problems using it in social sciences:

2. A control over experimentation is impossible in case of real-life events (being born in a poor family, being male, being migrant etc.).

3. Control over experimentation might be possible for certain projects (credit, extension, education), but it is often not feasible.

4. Even in case of random assignment groups can learn from each other => so-called spill-over effects

=> That's why quasi-experimental designs are still used in social sciences!

# Quasi-experimental research design

1. Similar structure as true experiments (treatment and control group).

2. Researcher lacks control of the treatment

=> Who the treatment receives does not rely on random assignment, but:

- on self-selection or
- administrative decisions

=> Treatment and control group differ in many variables influencing the outcome

=> Direct comparison is not possible anymore

- compare the groups according to baseline survey
- apply regression models to control for self-selection

# Ex-post facto design

Many independent variables tend to be natural or life experiences

=> they cannot be controlled directly by the researcher

=> they only can be observed by the researcher after the event/fact

=> use ex post facto research designs.

Regression analysis to make causal inferences or predictions (In contrary, experimental designs may not need regression analysis at all for causal inference).

# Stages in the research process

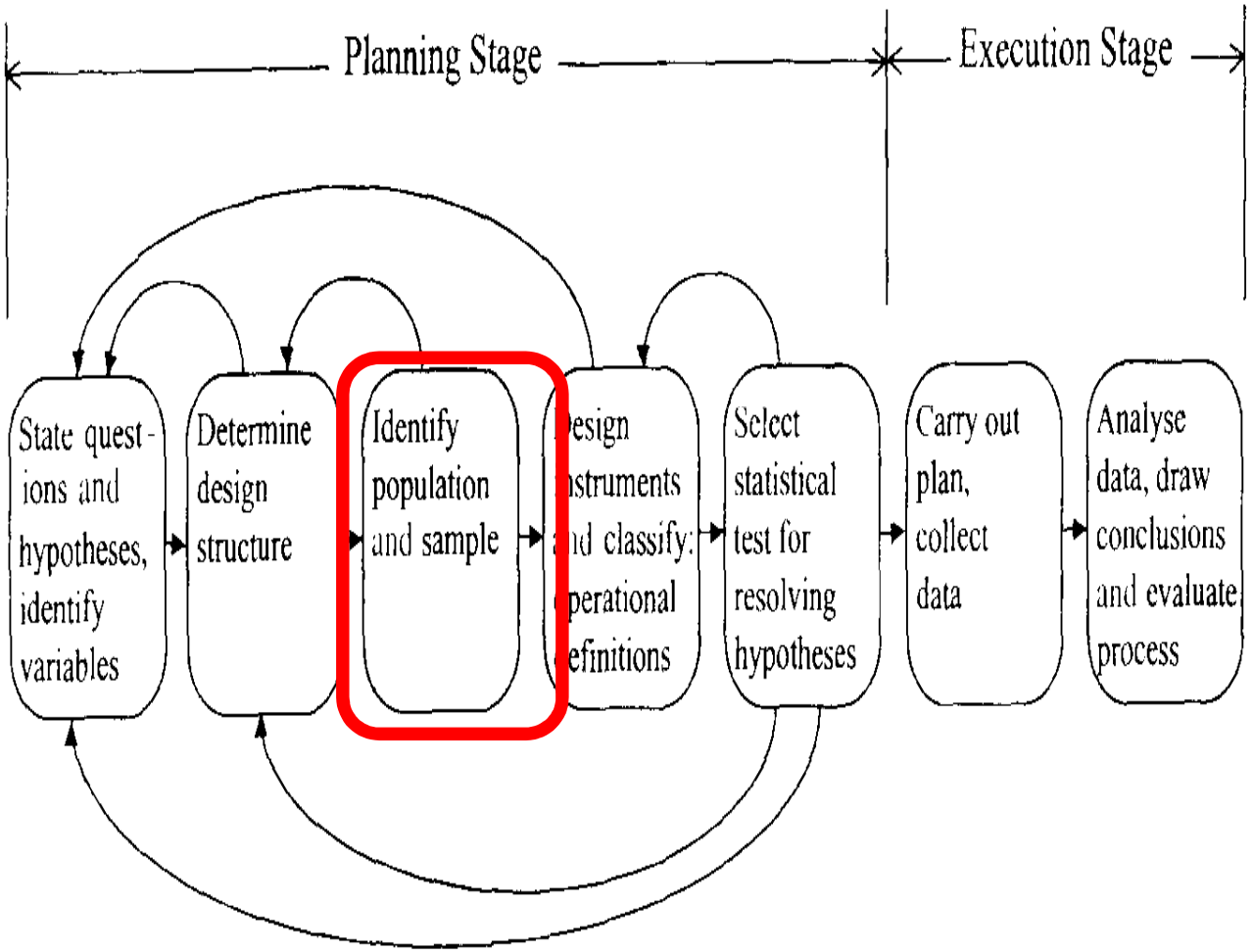



*FIGURE 2.1*
Stages of designing and carrying out a study, including iterations for modifications and improvements during planning

Source: Black, 1999

# Why take samples?

Major reasons:

1. A survey is less costly and time demanding than a census

2. More time can be spend on interviewer training and questionnaire development

   => increases data quality

3. Data processing and cleaning is much faster

4. We can infer from the sample to the population

=> Censuses are not necessary!

# Terms in sampling

Element (or observation unit)

=> the unit about which information is sought (wheat farmers).

Population (or universe)

=> is the aggregate of all the elements defined in terms of:

- elements (wheat farmers)
- extent (South Nyanza in Kenya)
- time (survey at the end of the growing season 1999/2000)

Sampling unit

=> elements available for selection at some stage in the sampling process

Sampling frame

=> list of sampling units available for selection

# The sampling process

Step 1: Define the population in terms of elements, extent, time, and sampling units.

Step 2: Identify the sampling frame

Step 3: Determine the sample size

Step 4: Select a sampling procedure

Step 6: Select the sample

# Types of sampling procedures

Two major types:

1. Non-probability sample

   => case study/exploratory research

   => not for causal research or for making inferences

   Procedures for selecting samples:
   - snowball sampling (respondent proposes next one)
   - purposive sampling (choice of researcher)
   - quota sampling (the first 30 wheat farmers that one encounters)

2. Probability sample

   There are four types:
   - Simple random
   - Stratified random
   - Cluster
   - Stage

# Simple random sampling

There is only one sampling unit (for example all wheat 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,
- a new census solely done for the purpose of the research.

Golden rule: Each and every element of the population has the same probability of being in the sample.

How to draw a random sample:

1. Using random selection features in EXCEL or SPSS

Excel: RAND ()*(100-1)+1

SPSS: data - select cases – random sample

# Simple random sampling (2)

How to draw a random sample (continued):

2. Mix equal-size papers with household names, and draw them, say from a hat (or let the village head draw them)

3. Systematic sampling

Start from a randomly chosen point on your list

=> select the sample in equal distance from the previous point chosen

=> going through the WHOLE list.

Example: sample of 10 from a list of 150 households

1. Determine the interval (150 divided by 10)
2. Randomly select a number between 1 and 15 => first household selected
3. Every 15th one on the list is selected

Potential danger: hidden pattern in the list.

# Stratified random sampling

Stratified random sampling can improve the accuracy of estimation, when there are at least two sampling units (for example certain type of farmers, lowland vs. upland wheat farmers), which:

- vary a lot between each other
- are homogeneous within

Like in simple random sampling, one needs a complete list of population members differentiated into the different strata.

Strategies for sampling:

1. Proportional sampling

   Assume 60% lowland wheat farmers and a sample size of 100
   => randomly select 60 out of all the lowland wheat farmers

   => randomly select 40 out of all the upland wheat farmers

# Stratified random sampling (2)

Strategies for sampling:

2. Disproportional sampling

We randomly select more units of a certain strata that is not very frequent in the population.

Assume that we are particularly interested in female wheat farmers, but only 10% of the wheat farmers are female:

=> in a sample of one hundred we would expect just 10 female wheat farmers

=> we can now easily select 20 female wheat farmers

Is inference from the sample to the population possible?

# Stratified random sampling (3)

Strategies for sampling:

2. Disproportional sampling

We are able to randomly select more of a certain strata that is not very frequent in the population

Assume that we are particularly interested in female wheat farmers, but only 10% of the wheat farmers are female

=> in a sample of one hundred we would expect just 10 female wheat farmers

=> we can now easily select 20 female wheat farmers

Attention! Because the sample size in the strata is NOT proportionate to the relative frequency of the strata in the population we must use so-called sampling weights in the data analysis.

# How to calculate sampling weights?

$$W_i = \frac{\frac{n_i}{N}}{\frac{s_i}{S}}$$

Wi: Weight strata I

ni: Number of elements in strata I

N: Total number of elements

si: Number of elements sampled from strata I

S: Size of total sample

Example:

100 villages; 10 have no road access

We randomly select 8 villages of which 50% have no road access

# How to calculate sampling weights?

$$W_1 = \frac{\frac{10}{100}}{\frac{4}{8}} = 0{,}2 \qquad W_2 = \frac{\frac{90}{100}}{\frac{4}{8}} = 1{,}8$$

For sampling weights it must always hold true, that:

sum of sampling weights for all elements selected in the sample

=

sample size

# Cluster sampling procedure

Problem: often lists of the population members do not exists and are costly to obtain

=> Take random samples of successive clusters of subjects

Example:

Your population consists of 16000 households living in 80 villages.

There exist no lists of households, but a list of all the villages including the number of residents.

1. Randomly select villages
2. Go to the villages selected and obtain a list from village headman
3. Take a random sample of households in each village

In case of clusters of unequal size the selection procedure has to be changed to guarantee that each household has the same chance of selection.

# Cluster sampling procedure (2)

Example: 3 villages with 20, 30, and 50 households
=> want to select 20 households in 2 villages

1. Probability Proportional to Size (PPS)

- In case of simple random sampling households would not have the same probability of being selected.

=> Create number between 1 and 100 to select the cluster, before randomly selecting the households.

| Village | Size | Cumulative size | Sampling Interval | Unit selected | Cluster selected |
|---|---|---|---|---|---|
| 1 | 20 | 20 | 1-20 | | |
| 2 | 30 | 50 | 21-50 | 37 | X |
| 3 | 50 | 100 | 51-100 | 88 | X |

# Cluster sampling procedure (3)

Example: 3 villages with 20, 30, and 50 households
=> want to select 20 households in 2 villages

2. Equal Probability of Selection Method

- Select 2 villages using simple random sampling
- Assume that we selected village 2 and 3
- Sample size in village 2: 30/80 * 20 = 8
- Sample size in village 3: 50/80 * 20 = 13

Same probability of being selected?

# Stage sampling procedure

… is a combination of cluster sampling and strata sampling

Example:

1. Villages are grouped in two strata (with or without road access).
2. In each strata, a random selection of villages is taken.
3. In each of these villages, households are randomly selected.

Advantage:

It combines the advantages of cluster and strata sampling.

Disadvantages:

- It is a complex undertaking.
- The strata and clusters must be carefully defined.
- Available econometric modeling options is lower.

# Stage sampling procedure: Example

Collaborative Research Centre ‘Stability of Rainforest margins’

Research area: 722,000 ha, 136,707 people living in 117 villages

Objectives for sampling frame:

1. Random selection of observation units (villages, households)
   => infer to population
2. Focus of research is the vicinity of Lore-Lindu National park
   => stratified sampling to select more observations close to park
3. Reduction of costs for transport and logistics
   => cluster sampling

Steps in sampling:

- Stratification of villages according to distance to national park
- Random selection of villages within each stratum
- Random selection of households within the selected villages

# How to obtain a sampling frame?

**Method 1: Using secondary data**

When using secondary data it is important to judge its quality:

- When was the census made? Much fluctuations? Fairly accurate?

If we cannot answer these questions, we may do a census ourselves in a small village => compare it to the official numbers.

In general for developing countries:

- Information on villages is usually accurate.
- Data on households are almost always questionable.

=> Obtaining a list of households from the local list provided by the village head is often better than searching through piles of data in some ministry.

=> We may update this list with the heads knowledge.

# How to obtain a sampling frame?

**Method 2: Doing our own census**

1. Invite the village elders, teachers, religious leaders (all people that supposedly might know the residents of the village)

   => ask them for the names of the heads of the households

2. Undertake a census by walking from house to house

   => is a somewhat obtrusive undertaking

   => need permission and active support of the village leaders

In large villages it is often advisable to randomly choose certain sections.

# How to obtain a sampling frame?

**Method 3: Random sampling without a sampling frame**

Sampling without a sampling frame can be done, if method 1 and 2 are too time-consuming and costly.

There are various methods described in social science research books

=> one often used in field research in urban and rural areas of developing countries is the RANDOM WALK Method

=> developed by UNICEF/WHO for sampling households with pre-school children

# The random walk

1. Approximate the village or locality boundaries (draw a map).
2. Determine a central point and assess density of households.
3. Divide area into quarters (reflecting approximate 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).
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.
6. Replace non-response/drop-out households by sampling the very next household.

# Determining the sample size

The sample size required to obtain an estimate of a variable X within a certain probability of error cannot be calculated unless

=> one a priori knows the variability of variable X within the population

Unfortunately, we usually do not know the distributions of our variables of interest before the research.

Exceptions are published national data/prior surveys from which one can make best guesses of how the distribution may look like.

Because

- we do not know a priori the distributions
- we are usually concerned with many variables

=> the optimal sample size is almost never been calculated and used as a decision criteria for research surveys

# Determining the sample size: Some practical hints

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.

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.

3. The bigger the sample size, the better.

   BUT: Usually, with a larger sample size, so-called non-sampling errors increase, such as interviewer errors, non-response errors, and data processing errors.

# Determining the sample size: Some practical hints (2)

4. You need a lower sample size if your questionnaire is carefully pre-tested => interviewer and respondent errors are low

5. You need a lower sample size for the same precision if we employ stratified random sampling instead of simple random sampling.

6. Holding budget and time constant, we can increase our sample size by reducing the size 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 him.

# Determining the sample size: Some practical hints (3)

7. Descriptive studies that seek to document frequencies, means, and bivariate cross-tabulations may yield reasonable results with a sample size between 60 to 100.

8. Causal studies which employ multivariate regression models (OLS, Probit, or Logit) should have a minimum of 100.

9. Causal studies that wish to employ complex regression, such as two-stage and three-stage regression models, should have sample sizes above 200.

# Stages in the research process

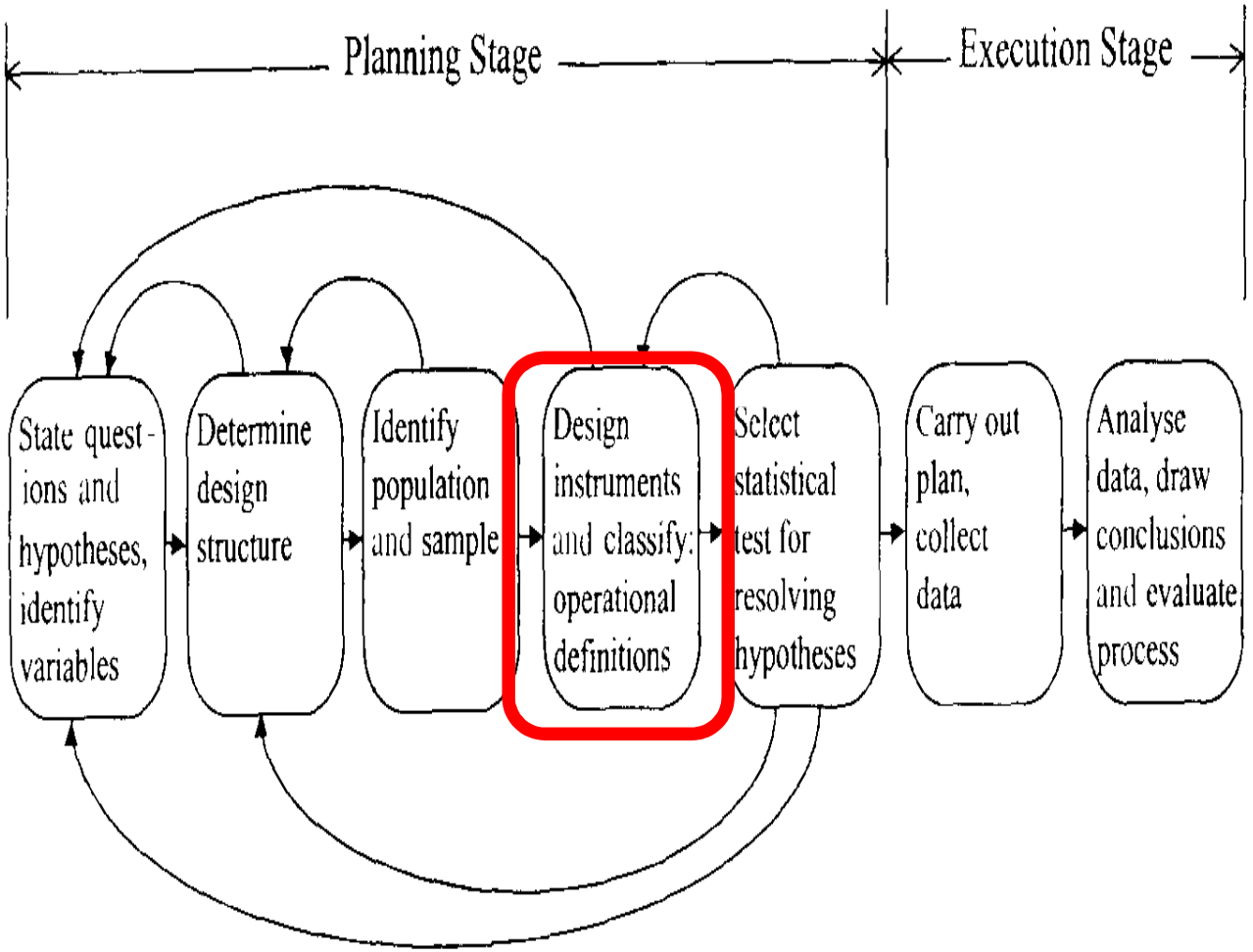



*FIGURE 2.1*
Stages of designing and carrying out a study, including iterations for modifications and improvements during planning

Source: Black, 1999

# Components of this stage

1. Identifying empirically operational measures for the variables in our conceptual framework

   => choose the measurement scale for the variables

2. Decide about mode of administration

   Postal questionnaire; Personal interview; Phone interview; Internet questionnaire

3. Designing the questionnaire

   - Deciding on structure of questionnaire
   - Lay-out of questionnaire
   - Phrasing of questions

4. Pre-testing the questionnaire in near-field conditions
5. Revising/finalizing the questionnaire

# Measurement scales of variables

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

=> These are often categories, such as areas, religion, ethnic groups, etc.

Note: Try to avoid to enter so-called string variables, like names of villages, into the spreadsheet if we wish to analyze the data.

=> Redefine them always into nominal variables.

Permissible statistic procedures:

Percentages, Mode, Chi-Square test

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

# Measurement scales of variables (2)

**2. Ordinal variables:**

Order of numerals in a hierarchical sense, e.g. 1 < 2 < 3

=> These are often variables measuring levels of achievement (education, salary ranks, farm size ranks, etc.)

Permissible statistic procedures: All test permissible for nominal variables, plus percentiles, median, and rank-order correlation.

**Example for coding education:**
**1= did not complete primary school, 2 = did not complete secondary school, 3 = secondary school education or higher**

# Measurement scales of variables (3)

**3. Interval variables:**

Equal differences between levels of measurement, i.e. (2 – 1 = 7 – 6).

=> These are often variables measuring attitudes, opinions, index numbers.

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 ...?'

Permissible statistic procedures: All tests permissible for nominal and ordinal variables, plus range, mean, standard deviation, and product-moment correlation

# Measurement scales of variables (4)

**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.

Note: It is important to use units of measurement that are known and normally used by the respondent.

Permissible statistic procedures: All test permissible for nominal, ordinal and interval variables, plus geometric and harmonic mean, and coefficient of variation.

# Questionnaire design

What is a questionnaire?

It is a standardized form for collecting data from respondents.

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.

A clear concept is needed (What information is needed? How will the data be analyzed?).

Good questionnaires are often developed in stages, involve local researchers and extensive pre-testing.

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

# Wording of questions

The words used in a question:

- should be familiar to the interviewer and respondent
- should correspond to local word usage and practices

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 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.

# Wording of questions (2)

Well-posed questions:

1. Use simple and clear words.
2. Avoid leading questions that bias the answer.

   *Was the harvest good last year?*

   *How was the harvest last year?*
3. Avoid implicit alternatives and assumptions.

   *Can you afford a telephone?*

   => carries the assumption that the respondent cannot afford it.

   *Do you have a telephone in your house for your own use?*
4. Avoid, if possible, estimates. However, many questions ask for estimates (e.g. yield, harvest, amount consumed)

   *How much rice [the basic food staple] did you consume in last year?*

   *How much rice did you consume during the last 7 days?*

# Wording of questions (3)

Well-posed questions (continued):

5. Avoid double barreled questions.

   *If you are neither a farmer nor a fisher, do you earn your living from trade?*

   *What is your main source of income?*

6. Consider the frame of reference.

   *Was any cattle stolen?*

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

7. Indicate to which respondents the question applies.

   *Do households sell wheat in the market?*

   *Do you sell wheat in the market?*

   Golden rule: Always try to ask the person directly to whom the question applies

=> Try to phrase the questions as simple and precise as possible!

# Types of questions

1. Open-ended questions

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

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

*What school education do you have?*

If answers are clearly defined and pre-known

=> use a pre-tested code that the interviewer can apply during the interview

# Types of questions (2)

2. Multiple choice questions

The interviewer reads the answers and the respondent has several options from which to choose.

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

3. Dichotomous question: The respondent answers with a “yes” or a “no”.

*Are you satisfied with the wheat extension service?*

# Coding of questions

**Interval/ratio data**

=> units are often monetary, weight, distance, or other straightforward measures

=> specify them in the question or use codes for the different units used

**Nominal (categorical) and ordinal data**

=> potential answers may be pre-coded

=> need to code the answers later for quantitative analysis.

However, for larger surveys which are not done personally by us, but by interviewers, the questionnaire should already contain pre-coded answers.

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

# Structure of the questionnaire

**Questions should logically flow**

=> Do not jump from household composition to income, and then back to education of household members.

**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.
- The most sensitive questions (questions that may offend the respondent)
  - we should try to avoid
  - if it cannot be avoided we put it at the end of the questionnaire.
- Questions regarding attitudes, preferences, subjective ratings of other people or institutions should come last.

  => They are almost always sensitive.

# Structure of the questionnaire (2)

**A questionnaire usually contains the following five sections:**

1. Identification data:
   date, name of household and village etc.
   => Information on the household (name or id) should appear on each and every page of the questionnaire.

2. Request for cooperation:
   Include a couple of sentences on the first page of the questionnaire that the interviewer is asked to read before the survey commences:
   - purpose of survey
   - request the cooperation of the respondent.

   The cooperation request should always state that the survey will not have any direct benefits for the respondent, and that all information is treated anonymously.

# Structure of the questionnaire (3)

**A questionnaire usually contains the following five sections:**

3. Instructions for interviewers:

   Put related instructions on the questionnaire right after or before certain 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. Classification data section:

   This information obtains the socio-economic characteristics of the respondent's household or village.

5. Information section:

   This is the main section of the questionnaire.

# Exercise: Developing a questionnaire

**Task 1: Develop a questionnaire for agricultural activities, so that a production function can be estimated.**

Example: paddy rice production in Timor Leste

**Table 6.3: Maximum likelihood parameter estimates of the Stochastic Frontier**

| Variable | Parameter | ML estimate[1] |
|---|---|---|
| Coefficients in the Stochastic Frontier production functions : Cobb-Douglas functional form | | |
| Constant | β0 | -0. 841 (0. 299) |
| Score | β01 | 0. 036 (0. 018)** |
| Matdumm | β02 | 0.104 (0. 165) |
| Material | β1 | 0.080 (0.045)* |
| Land | β2 | 0.246 (0.054)*** |
| Seed | β3 | 0.163 (0.039)*** |
| Labour | β4 | 0.365 (0.047)*** |
| **Returns – to – scale** | | **0.854 (0.071)** |

*(**)[***] Significant at the 10% (5%) [1%] level of error probability.

1. Values in parentheses are standard errors.

Source: Own data

Source: Poku-Blaeschke, 2010

# Exercise: Developing a questionnaire

**Task 1: Develop a questionnaire for agricultural activities, so that a production function can be estimated.**

Necessary information: yield and production inputs (land, labor, capital for seeds, fertilizer, pesticides etc.)

Example: Questionnaire to assess the livelihood and particularly the cultivation of paddy rice.

Exercise:

- Form groups of 2-3 people.
- Each group is developing the plot-specific questionnaire for one of the most important crops in the research area (grapes, wheat, maize etc.).
- Present your draft questionnaire

# Best practices in data entry

1. You can save much time in data analysis if you design your data entry template properly **before** you start entering the data.

2. Each case entered in a Spreadsheet begins with the so-called key variables, which uniquely identify the case.

3. Never aggregate information from the questionnaire before data entry.

   Example: Household rooster

   - do not define a variable for each member (AGE1, AGE2, etc)
   - Instead create a member-specific file (use key variables and only ONE variable for AGE).
   - Aggregate information to any desired level using software.
   - It is VERY cumbersome to break the information up again to lower levels.

# Best practices in data entry

4. We avoid big files (with more than 100 variables in one file)

   Break the data base up in several files, usually corresponding to the different modules of the questions.

5. There is no variable in a data file that does not have a variable label and value labels (for nominal or ordinal variable).

6. It is good practice to retain exactly the same coding system that is specified in the questionnaire (if possible).

7. Choose variable names that have some meaning, like age, gender, hhsize etc. (and not x23 or y24).

8. SPSS, MS Access, and other data entry programs, have checks for value ranges.

# Best practices in data cleaning

Entered data may contain at least three types of errors:

- Respondent error
- Interviewer error
- Data entry error

The following practices for data cleaning are frequently used:

1. Check for missing values

   If values are missing, then

   1. Check the original questionnaire(s)
   2. If the answer is also missing in the questionnaire, it can be an interviewer error.
   3. Consult the enumerator.

   If the enumerator can not remember that case, enter a missing value.

# Best practices in data cleaning

2. Check for wild codes

We want to get rid of all codes that do not exist for a particular variable.

In case of wild codes

=> check the questionnaire

=> ask the enumerator

3. Extreme case check (so-called outliers).

=> Checks for particularly high values

A household was found to hold land assets worth nearly $500,000.

=> too many zeros have been entered

# Best practices in data cleaning

4. Consistency Checks

=> Checks the logical patterns of answers

=> Compares more than one variable

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.

# Stages in the research process

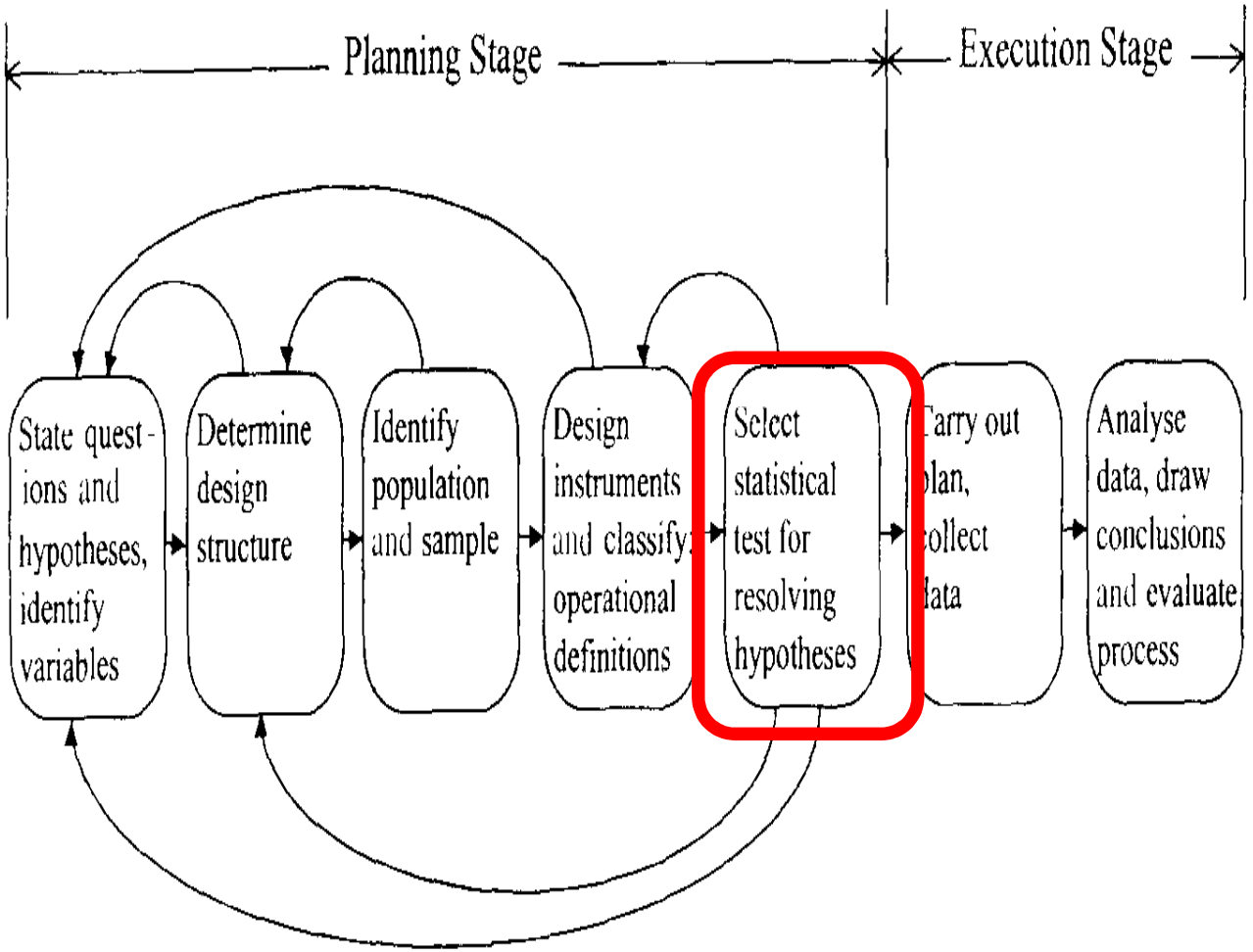



*FIGURE 2.1*
Stages of designing and carrying out a study, including iterations for modifications and improvements during planning

Source: Black, 1999

# Analyzing quantitative data 1

- One important domain for socio-economic research in conservation planning is the assessment of the costs of conservation.

- Most difficult task is to determine the opportunity costs of conservation:
  What is the potential economic benefits from using the park area for economic activities instead of protecting it?

- Such an assessment is commonly based on:
  - Gross margin analysis
  - Production function analysis

What is a gross margin?

# Excurse: Gross margin analysis

The gross margin (gm) of an activity is defined as:

gm = gross income – variable costs

- It refers to a particular farm activity
- It usually refers to one year or to one season

Gross income consists of the output obtained in physical, which is valued depending on its use:

Amount sold: => valued with the producer price
Amount home consumed: => valued with the consumer price
Amount used as fodder: => valued with the feed price
Amount used as seeds: => valued with seed price

Always use prices at the time of harvest => storing is another activity!

# Excurse: Gross margin analysis

Variable costs

- Vary with the quantity produced
  => they rise and fall with the quantity produced/the level of operation

- Often called operating costs

- Examples:
  - inputs for crop production: seeds, fertilizer, and pesticides
  - inputs for livestock production: feed, medicine, procurement of livestock
  - hired labor for a specific job
  - interest on working capital
  - costs for machinery

# Excurse: Gross margin analysis

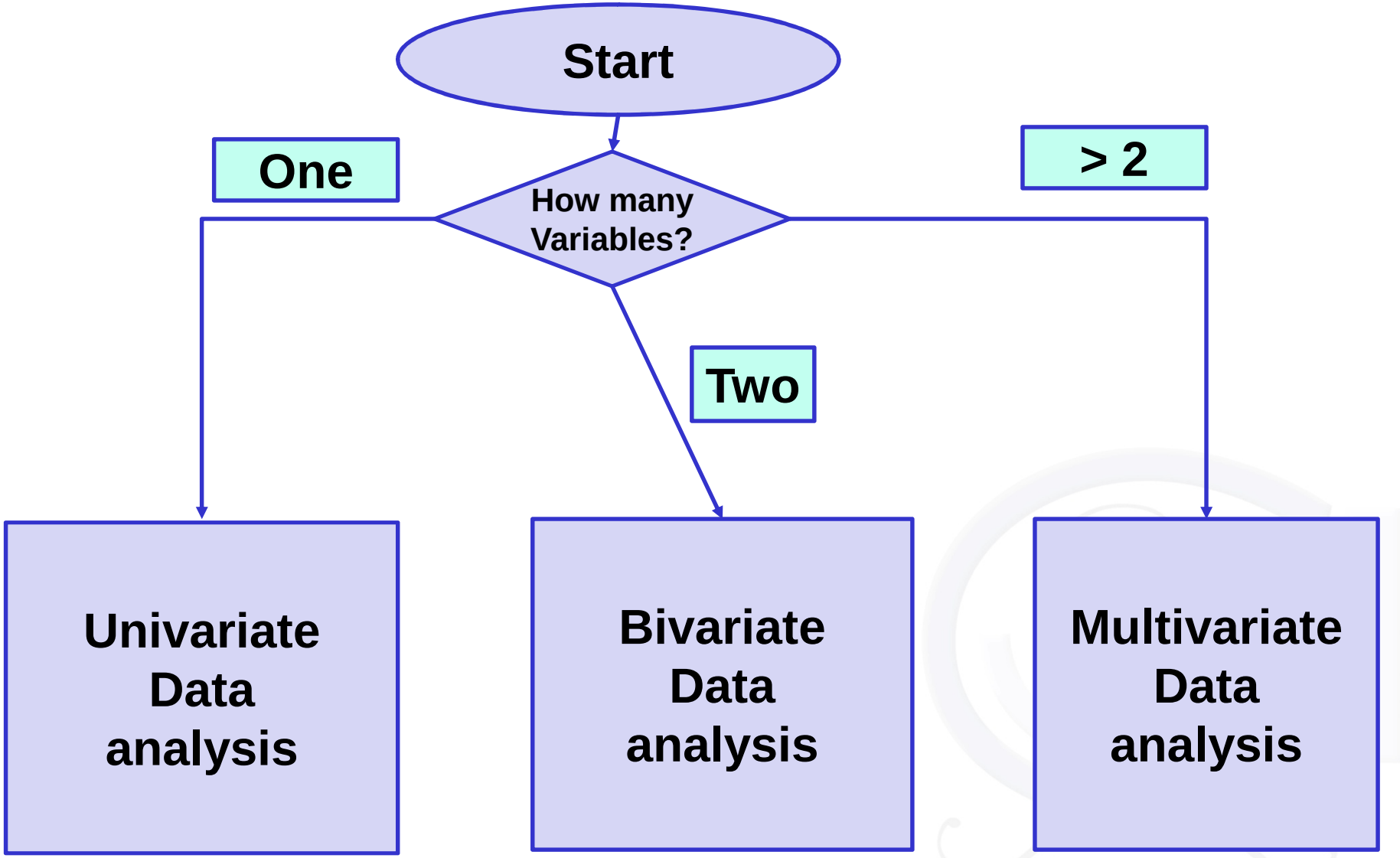
Example:
Gross margin calculation
for wheat production

| **Income** | Amount ton/ha | Price US$/ton | Amount US$/ha |
|---|---|---|---|
| wheat sold | 1 | 110 | 110 |
| wheat home consumed | 1 | 135 | 135 |
| wheat stored | 0.5 | 110 | 55 |
| **Total gross income** | | | **300** |

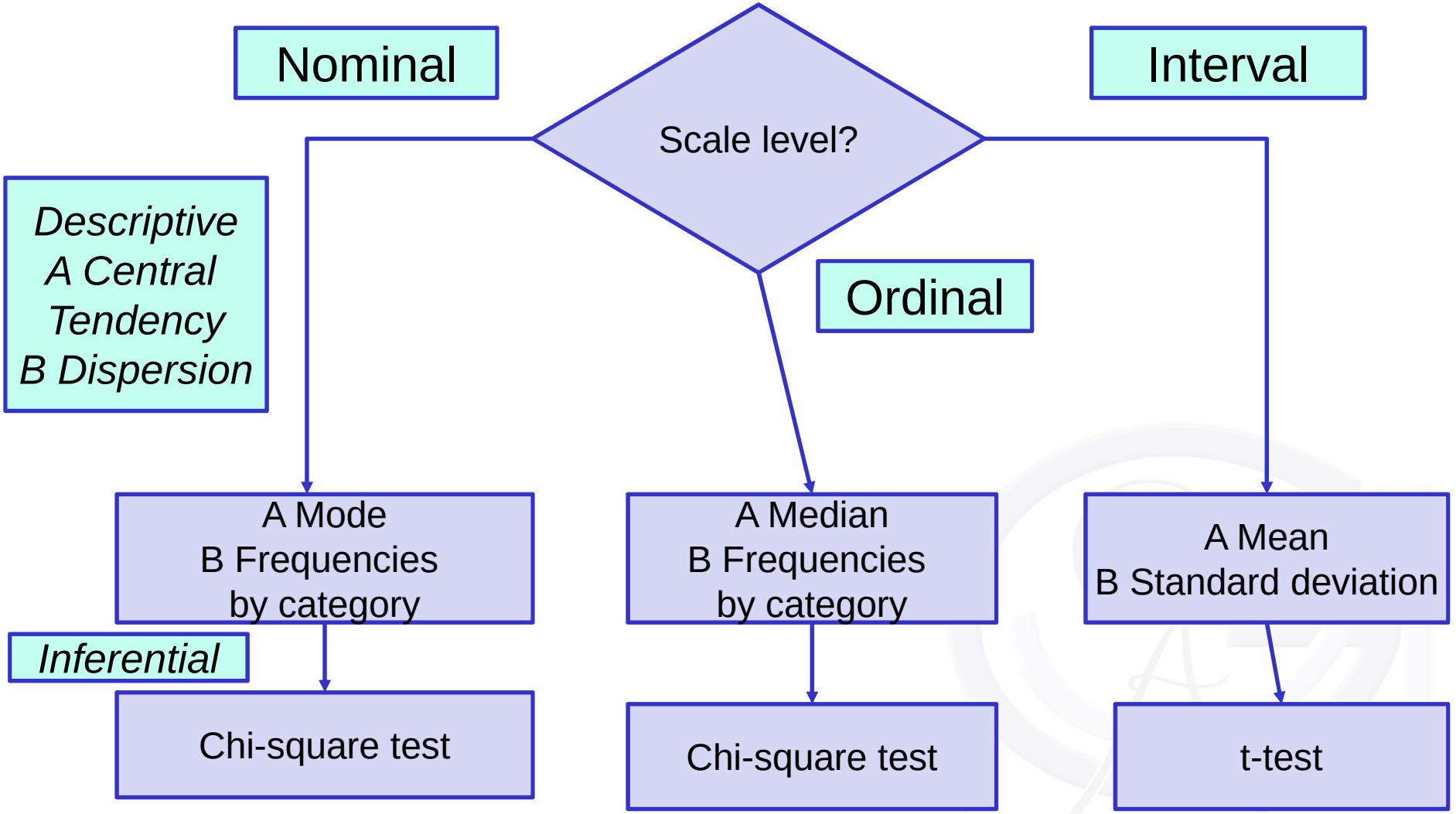
| **Variable costs** | Amount US$/ha |
|---|---|
| seeds | 10 |
| fertiliser | 25 |
| hired labour | 45 |
| repair/maintenance | 15 |
| fuel | 15 |
| pesticides | 11 |
| crop insurance | 7 |
| interest for working capital | 5 |
| **Total variable costs** | **133** |

**Gross margin/ha** **167**

# Overview of data analysis techniques



# Univariate data analysis



# Univariate data analysis

## Nominal data: Chi-square test

Compares the observed distribution in a sample with an a priori expected distribution of a population.

$H_0$: There is no difference between the two distributions.

Self-assessment of household into 3 poverty categories (N=500):

| | Observed N | Expected number | Residual |
|---|---|---|---|
| Very poor | 63 | 166,7 | -103,7 |
| Poor | 387 | 166,7 | 220,3 |
| Better-off | 50 | 166,7 | -116,7 |

| | | |
|---|---|---|
| Statistik für Test:: | | |
| Chi-Square Value: | 437,428 | |
| Significance Level/Probability of error | | ,000 |

=> We can reject $H_0$ with a very low probability of error.

# Univariate data analysis

## Interval data: t-test

Compare the sample mean with a hypothesised population mean

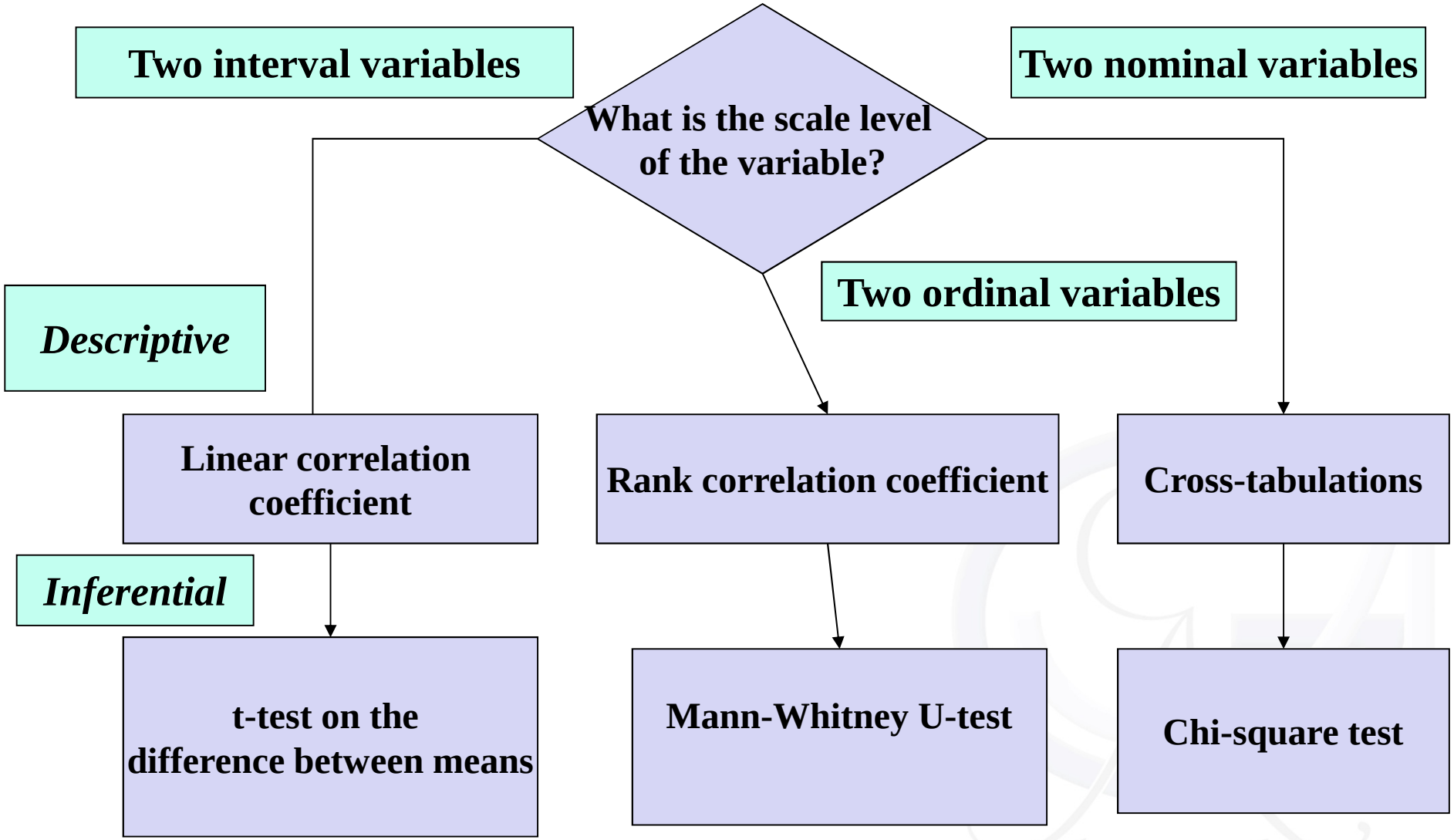
$H_0$: There is no difference

*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.*

$H_0$: The nitrogen level is 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: the sample does not support
our hypothesis that farmers apply 55 kg of nitrogen per hectare.

# Bivariate data analysis



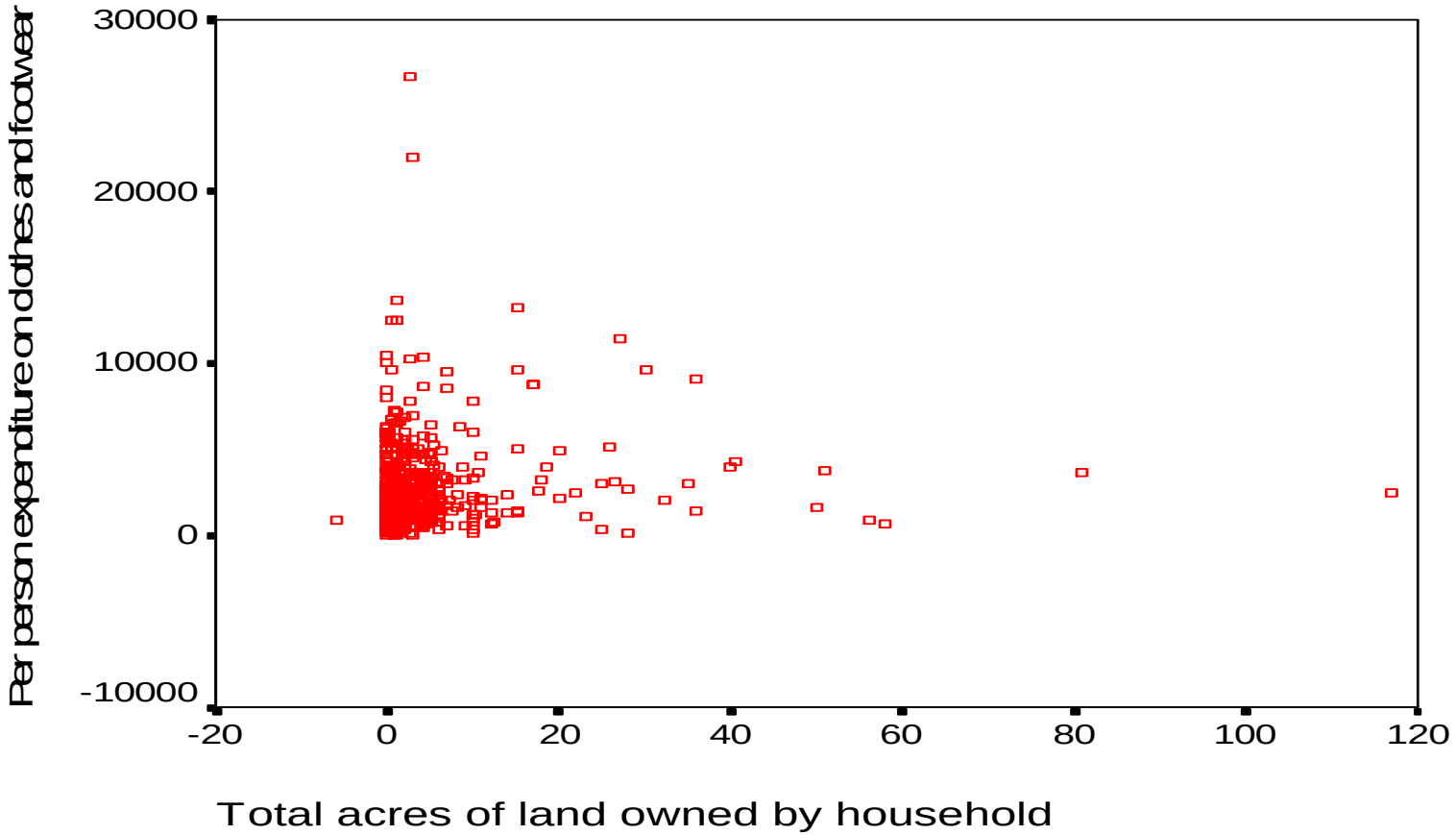
# Bivariate data analysis

**(1) Linear correlation coefficient**

Degree of association between two interval variables
(Note this is no indication for causality)

If the coefficient of correlation is:
0 => no correlation between X and Y.
> 0 => positive correlation
< 0 => negative correlation
> 0.8 very strong
0.4 – 0.8 moderate
Otherwise weak
Visualisation: scatter plot

# Bivariate data analysis




**Coefficient of correlation: 0,065**
**Probability of error: 14.5 %**

# Bivariate data analysis

## Nominal data: Chi-square test

Used in so-called cross-tabulations to identify relationships between two variables.
$H_0$: There is no relation between the two variables.

Example: Is their a relationship between the source of electricity and the condition of the house?

| | Condition of House | | |
|---|---|---|---|
| | Needs repairs | In good condition | Total |
| No electricity | 97 | 262 | 359 |
| Uses electricity | 11 | 130 | 141 |
| Total | 108 | 392 | 500 |

Chi-Square Value: 22.08 → Significance level: 0.000

# Bivariate data analysis

**Interval data: t-test**

Compares the mean of two sub-samples.

$H_0$: The mean age in the two groups is not different.

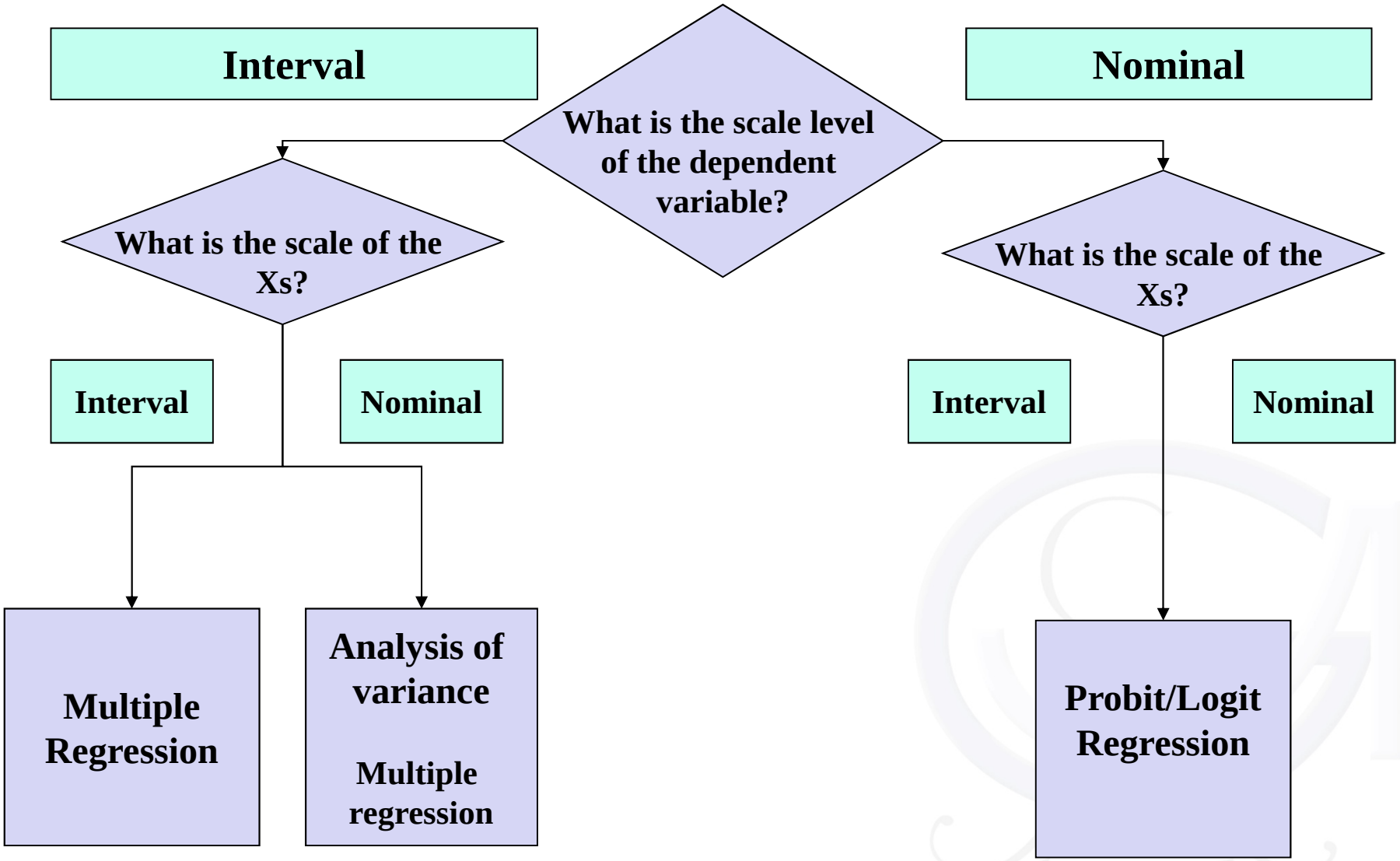
Example: Do members of a credit group have higher or lower clothing expenditures than non-members?

$H_0$: There is no difference in clothing expenditures between the two groups.

| | Member | N | Mean | Standard deviation |
|---|---|---|---|---|
| Clothing expenditure | No | 300 | 2383 | 2141 |
| | Yes | 200 | 3277 | 3131 |

T-Value: -3.526; Significance level: 0.000

# Multivariate data analysis



# Multivariate data analysis

Analysis of Variance (ANOVA)

We use this to test the null hypothesis that several (i.e. more than two) population means are equal.

One-way ANOVA

=> one variable is used for classification into groups

Factorial ANOVA

=> several variables are used for classification into groups

For example:

means of several ethnic groups further differentiated by marital status

# Multivariate data analysis

**Generic regression model with two independent variables:**

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

$Y_i$ : Dependent variable for case i =1 to N where N = number of cases

$ß_0$: Intercept (point on the y-axis at which all dependent variables assume the value zero)

$ß_1$, $ß_2$:Regression coefficient

$e_i$: Residual (unexplained variance)

Note: There are a number of tests available that need to be performed in order to test whether the model specification is appropriate. For an introduction, see for example, Kennedy: A Guide to Econometrics.

# Multivariate data analysis

**Output of a regression model estimated with ordinary least squares (OLS)**

| | B | T | Significance level |
|---|---|---|---|
| Constant | 5067 | 9,048 | ,000 |
| Average age of adults | -33 | -2,515 | ,012 |
| Household size | -316 | -5,893 | ,000 |
| Per adult value of land holdings | 2E-03 | 6,116 | ,000 |

Dependent variable Y: Per person expenditure on clothes and footwear

R-Squared: 0.12 → 12 % of variance in Y is explained by the Xs