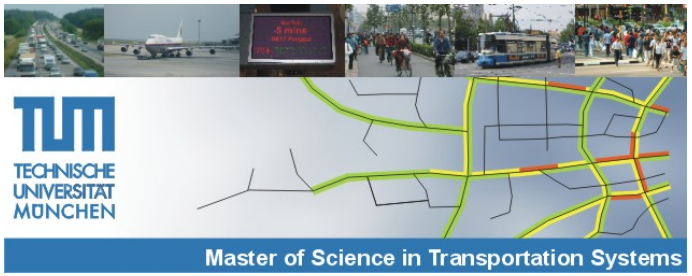
M.Sc. in ‚Transportation Systems‘



Seminar
“Hypothesis driven data analysis”

Tina Gehlert
Technische Universität München, mobil.TUM
Tina.Gehlert@gmx.de

Munich, 02/02/2012

Who I am …

Background

- 2002 M.Sc. in Psychology
- 2008 PhD in Traffic and Transportation Psychology

Employment

- Since 2008 Accident Research Department of the German Insurance Association, Senior researcher and project manager
- Since 2009 TU Munich, Guest researcher and guest lecturer

Research interests

- Psychological foundations of travel behaviour
- User reactions to road pricing (public acceptability, travel behaviour)
- Traffic norms and values e.g. traffic rule violations (red light running, speeding, drunk driving)

Contact: Tina.Gehlert@gmx.de

Plan for the 2-days Seminar

Learning objectives

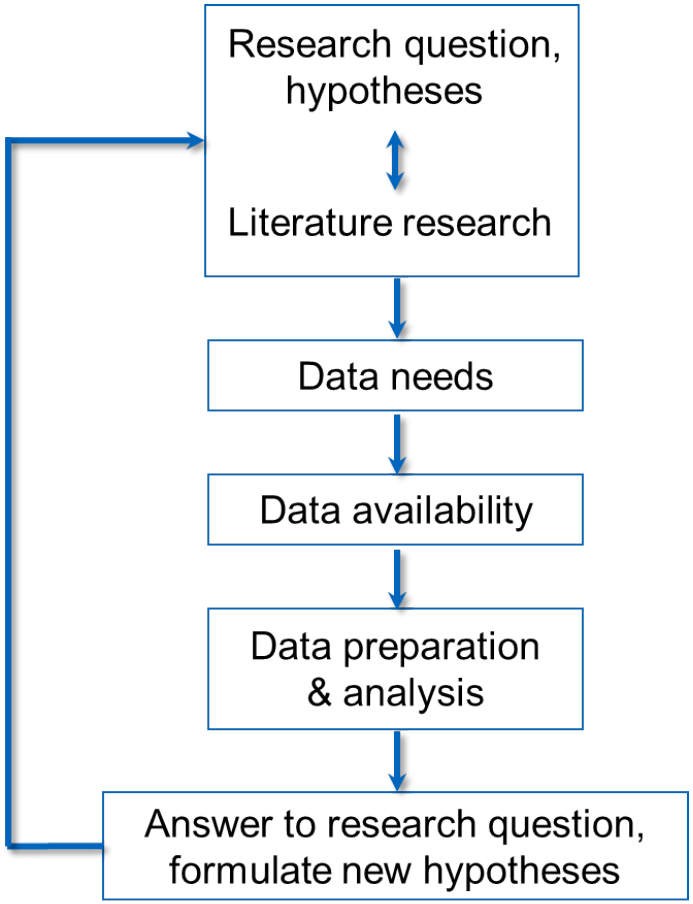
1. Develop an overall understanding of hypothesis-driven data analysis
2. Get to know the relevant steps along the empirical research cycle
3. Be able to conduct a hypothesis-driven data analysis

Learning tools

- Homework exercise
- Combination of theoretical and practical parts
- Case study with real life data (German Travel survey, MID 2008)

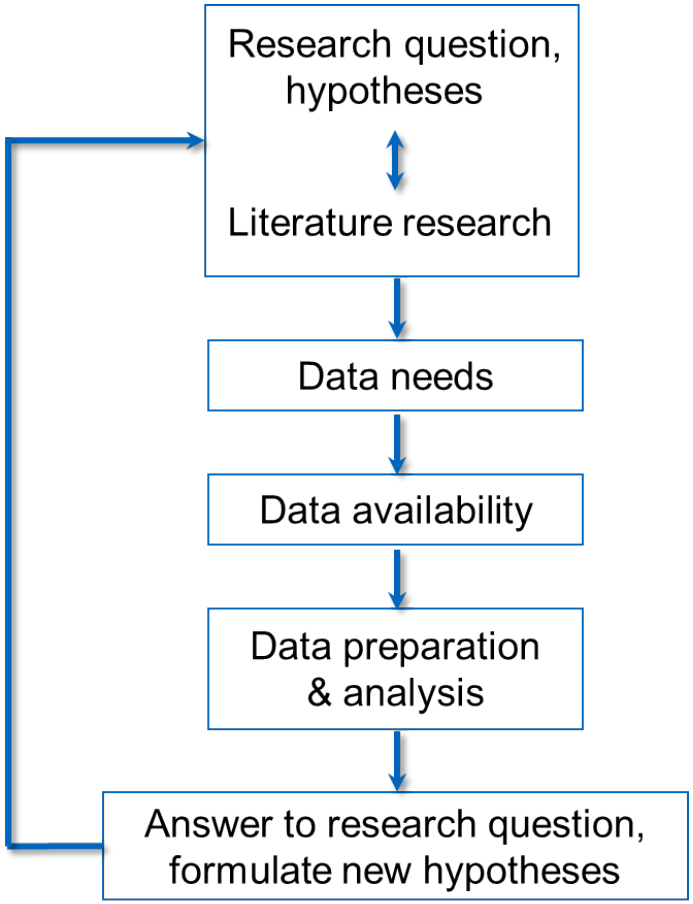
Lecture Plan – Thursday 2nd Feb 2012

1. Session:
   Repetition R (Gerike)
2. Session:
   Research question; Data needs & availability (Gehlert)
3. & 4. Session:
   Data preparation & analysis
   (Gerike / Gehlert)



Lecture Plan – Friday 3rd Feb 2012

1. Session:
   Data preparation & analysis (Gerike)
2. Session:
   Inferential statistics (Gehlert)
3. & 4. Session:
   Inferential statistics
   Discussion
   (Gehlert)



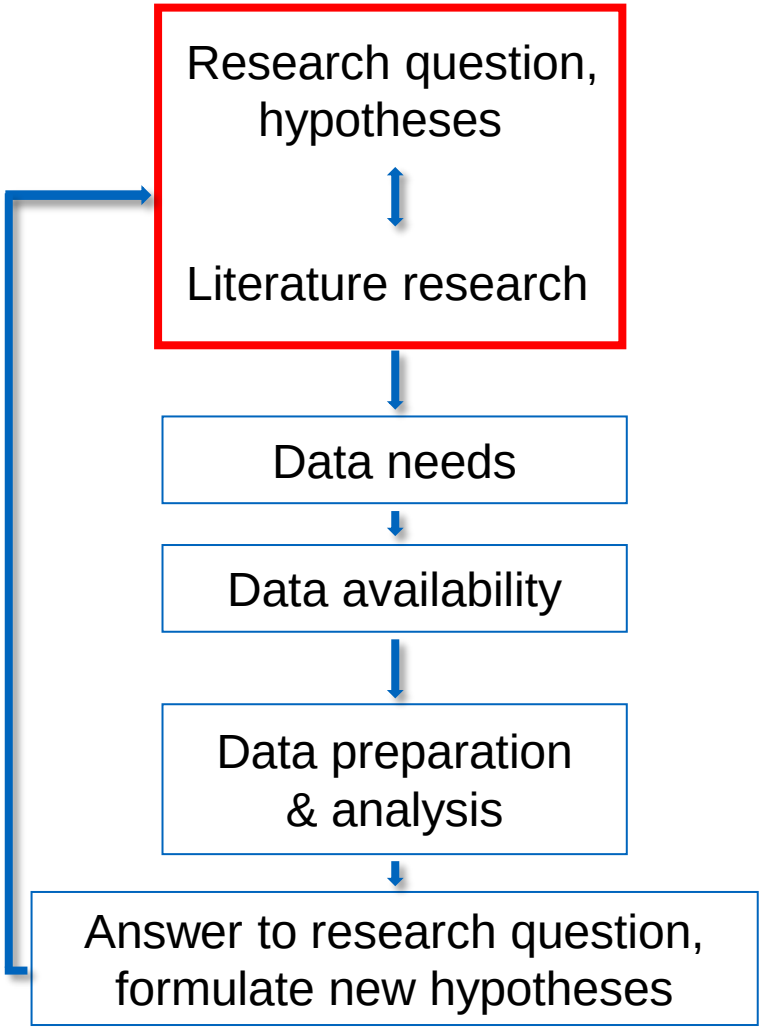
Lecture Plan

Step 1:

- Research questions and hypotheses

Learning objectives

- What is a research question?
- How is a hypothesis be formulated



Exercise 1 „Research questions & Hypotheses“

- Read the literature provided.
- Identify the 3 most important indicators of individual travel behaviour.
- Extract relevant factors (max. 3) that influence individual travel behaviour.
- Describe their impact on relevant indicators of travel behaviour.
- Identify and write down three questions for further research concerning these factors and their impact on individual travel behaviour.
- Formulate your expected answer to that question.

“Research is an organized and systematic way of finding answers to questions …”

A research question …

- provides the conceptual foundation of the data analysis
- focuses the data analysis, I.e. ensures that you choose the appropriate data source, variables and statistical tests

- States a *relationship* between two or more variables
- phrases the relationship in terms of a question.

Types of research questions (1)

There are three basic types of research questions:

Descriptive research questions:

- describes what is going on or what exists.
  - Does X exist?
  - What is X like? What are it‘s characteristics?
  - What are the components that make up X?

Example:

- Does a constant travel time budget exist?
- ….

Types of research questions (2)

Relational research questions:

- Looks at the relationships between two or more variables.
  - Is there an association between X and Y?
  - Is Group X different from Group Y?

Example:

- Is accessibility of transport modes related to individual travel behaviour?
- **How does accessibility of transport modes relate to individual travel behaviour?**
- **…**

Types of research questions (3)

Causal research questions:

- Looks at the relationships between two or more variables.
  - Does X cause, lead to, or prevent changes in Y?
  - Does X cause more change in Y than does Z?
  - Does X cause more change in Y than does Z under certain conditions but not under other conditions?

Examples:

- Does an extension of public infrastructure lead to a higher PT modal split?
- Does continuing increases in income raise total travel time expenditures per person per day?
- Do people with permanent car availability use more public transport than two decades ago?

Prerequisites of causality

1. A relation between DV and IV
2. The Independent Variable precedes the Depended variable in time
3. Other possible explanations are ruled out or controlled for.

Sources for research questions

- Literature in your specific field
- Practical problems in the field
- Theory
- Your own interest
- …

Hypotheses

- are expected answers to your research questions
- Should be:
  - specific,
  - testable,
  - relevant to the research question and
  - as simple as possible (Occam's razor)
- Include operational definitions of DV and IV and their relation
- Can be:
  - *one-tailed*
    (e.g., Men have a higher travel times than woman) or
  - *two-tailed*
    (e.g., Men and woman differ in their travel times).

Hypotheses: Example

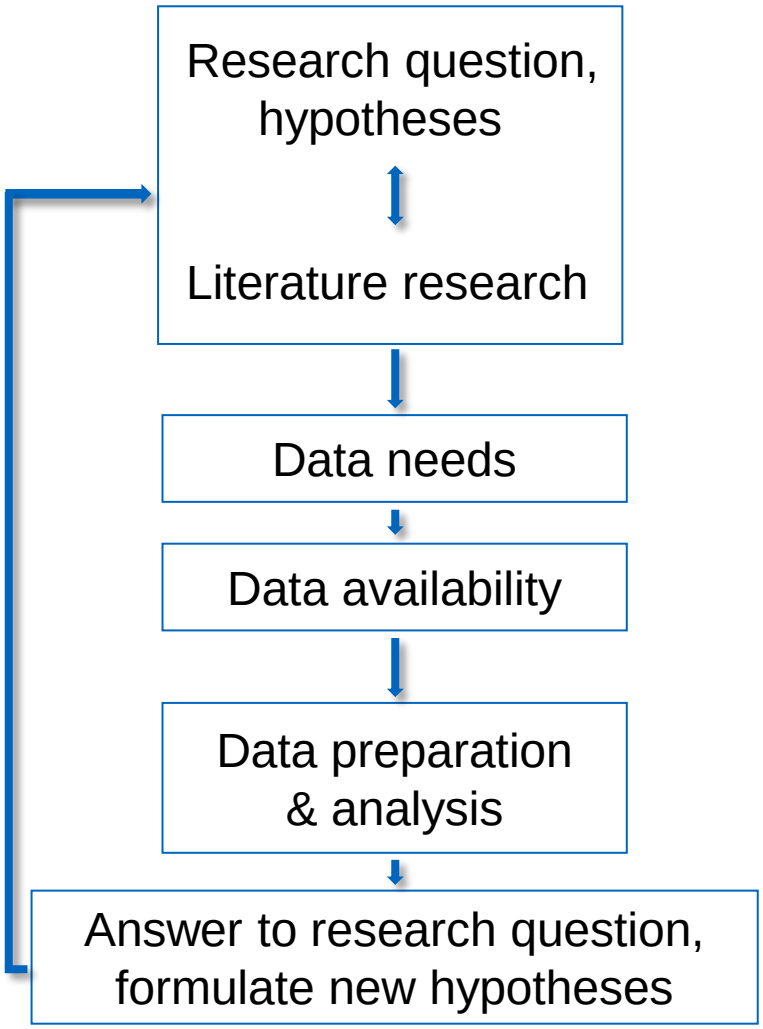
RQ: How does accessibility of travel modes relate to individual travel behaviour?

Accessibility	Travel behaviour
Availability of car / bicycle (yes / no)	No. of trips
Individual rating of accessibility	Trip distance
Frequency of mode use	Trip duration
…	…

CASE STUDY

RQ: How does accessibility relate to individual travel behaviour?

- H 1: The more often people use the bicycle the better the accessibility by bicycle is assessed for all trip purposes.
- H 2: People > 65 years old have less access to private cars and us them less often.
- H 3: Car availability relates to the number of trips, trip distance and trip time by car.



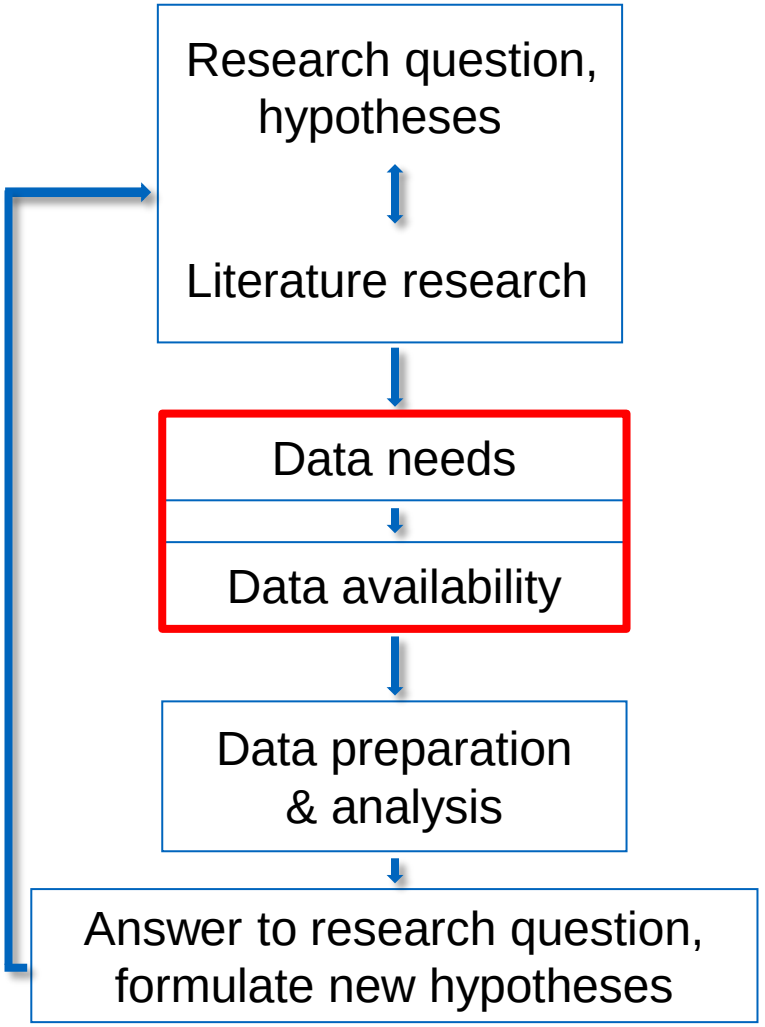
Lecture Plan

Step 2:

- Data needs & availability

Learning objectives

- How to derive data needs
- Where and how to find data



Data needs

RQ & Hypotheses	Data needs
Type of research question	Research design
Variables and its operational definitions	Independent and dependent variables
Population	Sample characteristics, sample sizes
Geographical area	Sample characterics / spatial information (e.g. geocoding)
Unit of analysis	Person / Trip / Day / Car
…	…

CASE STUDY

RQ: How does accessibility relate to individual travel behaviour?

- H 1: The more often people use the bicycle the better the accessibility by bicycle is assessed for all trip purposes.
- H 2: People > 65 years old have less access to private cars and us them less often.
- H 3: Car availability relates to the number of trips, trip distance and trip time by car.

Exercise 2 „Data needs and availability“

- Get familiar with the documentation the dataset MID 2008.
- Identify and write down for the first hypothesis (H1 variables from MID 2008 that could be used for the analysis

H 1: The more often people use the bicycle the better the accessibility by bicycle is assessed for all trip purposes.

Data needs – Case study

Research Question & Hypotheses H1 to H3	Data needs
RQ type: Relational	Cross sectional design
Variables and operational definitions IV: Accessibility DV: Travel behaviour	IV: car availability, distance to next PT stop, assessment of accessibility DV: No. trips, trip distance, trip time
Population of all travel modes	Sample: all travel modes , especially bicyclists, car users
Geographical area	No restrictions
Unit of analysis	Person
Sample size	Reasonably high

H1 (selected variables)

H 1: The more often people use the bicycle the better the accessibility by bicycle is assessed for all trip purposes.

Data needs for H3	Person data	Trip Data	Household data	Car data
IV bicycle usage	p033: Frequency of bicycle use		-	-
DV Accessibility by bike: subjective assessment	p0412_1: Accessibility of workplace by bicycle p0412_2: Accessibility of apprenticeship location by bicycle p0412_3: Accessibility of school by bicycle p0412_4: Accessibility of shops |businesses (daily shopping) by bicycle		-	-
Sample description age, gender	hp_sex: Gender (missing data supplemented from household interview) hp_alter: Age (missing data supplemented from household interview) hp_altg3: Age group (variable 3)			-

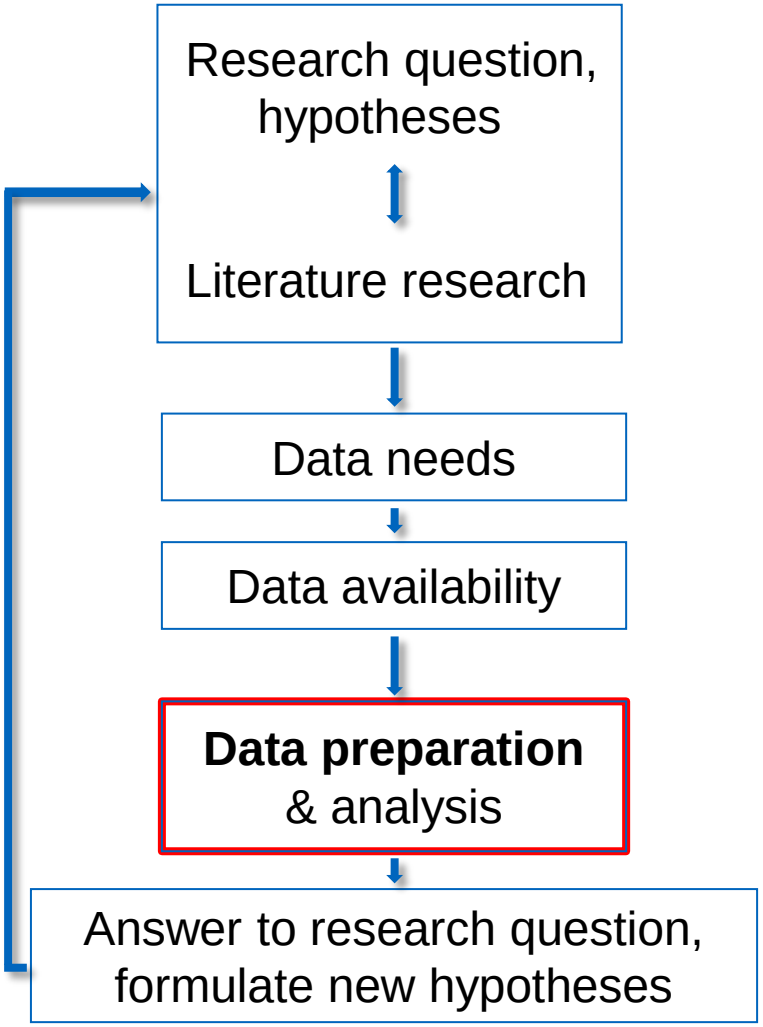
Lecture Plan

Step 3:

- **Data preparation** & analysis

Learning objectives

- What are the various tasks of data preparation?
- What are steps of data analysis?



Data preparation

Aims

- to verify the dataset (i.e. check for correspondence between data set and documentation)
- to get familiar with the data set
- to check for data entry characteristics and possible mistakes
- to prepare the data for subsequent analyses

Data preparation

Tasks

1. Verifying the data set
2. Aggregate data
3. Calculate new variables
4. Combine data sets
5. Choose subsamples

➢ At the end there will be one (or more) processed data file (or many) with which you can do your subsequent analyses!

Data preparation

Task 1: Verifying data sets

- Range checks (e.g. subjective asssessment of accessibility)
- Filter checks (e.g. accessibility of workplace by bicycle by hp_beruf)
- Consistency checks (e.g. different age variables)
- Descriptive statistics (e.g. subjective asssessment of accessibility (by car)

Task 2: Aggregate data

- Data is expressed in a summary form
- Usually done for trip data
- (e.g. MID 2008 Person data file: variable wege1: Reported trips (from trip survey))

Data preparation

Task 3: Calculate new variables

- Create new indices
- Recreate indices if different conceptualization

Task 4: Combine data sources

- Combining data from internal sources e.g. trip diaries, household interviews… etc.
- (e.g. MID2008 Person data: variable hp_pkwfs: Car drivers license (from p061_3 entry in the household interview)
- unique primary keys are important!
- Adding data from external sources
- (e.g. MID2008 spatial variables BBSR types)

Task 5: Choose subsamples

- Filter relevant cases
- Never delete cases!

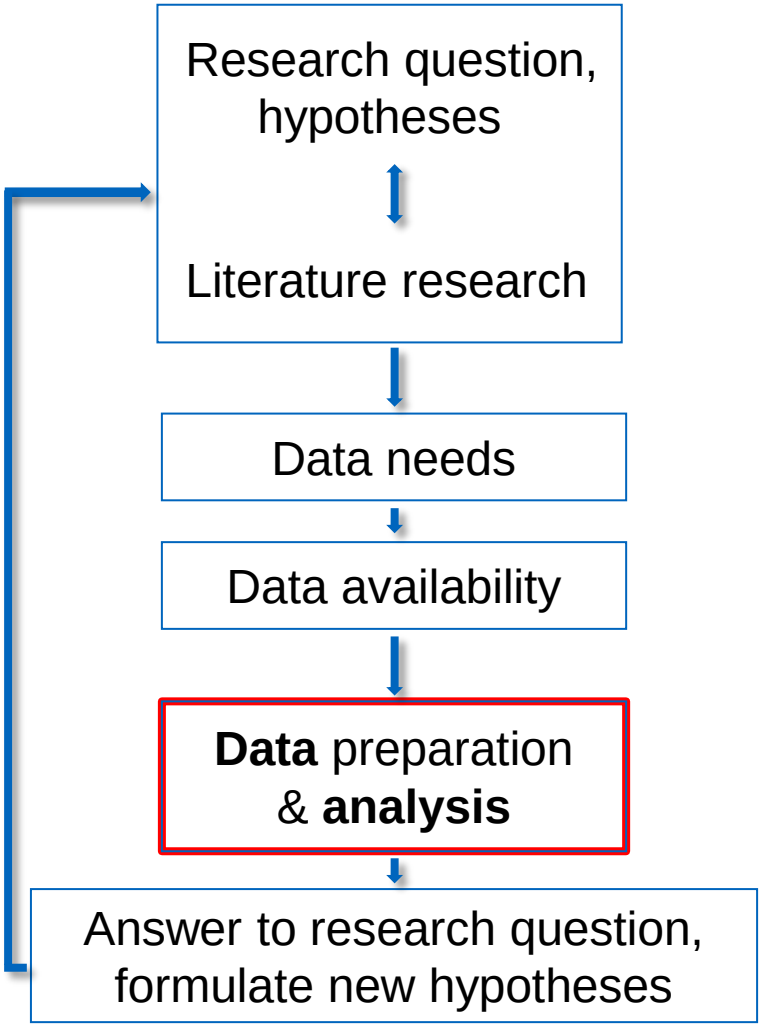
Lecture Plan

Step 3:

- **Data** preparation & **analysis**

Learning objectives

- What is descriptive and inferential analysis?
- Which indicators describe empirical data?



Data analysis

Two parts:

- **Descriptive statistics:**
  - provides a summary of the quantitative information about the data set
- **Inferential statistics**
  - Making decisions about a population based on the information contained in a random sample from that population

Descriptive Statistics

Central tendency

- is the tendency of quantitative data to cluster around some central value.
- Measures: mean, median, or mode (depending on the scale level)

Dispersion

- Describes the closeness with which the values surround the central value
- Measures: standard deviation or variance

Distribution: description of shape / symmetry

- important for the selection of statistical tests (normal distribution)

CASE STUDY

RQ: How does accessibility relate to individual travel behaviour?

- H 1: The more often people use the bicycle the better the accessibility by bicycle is assessed for all trip purposes.

Descriptive statistics (summary)

H 1: The more often people use the bicycle the better the accessibility by bicycle is assessed for all trip purposes.

Variables	Scale level	Central tendency	Dispersion	Distribution
Gender	Nominal	2 = women	Min = 1, Max = 2	48% men 52% women
Age	Metric	45.63 years	10.24 (SDDV)	No NV
Frequency of bicycle use	Ordinal	3 = 1-3 days per month	Min = 1 Max = 5	22.0% daily 23.2 % 1-3 d/w, 16.0% 1-3 d/m, 11.3% > m 27.5% never
Accessibility per bicycle workplace, shopping	Metric	2.44 = between 'good' and 'satisfactory'	1.29 (SDDV)	No NV

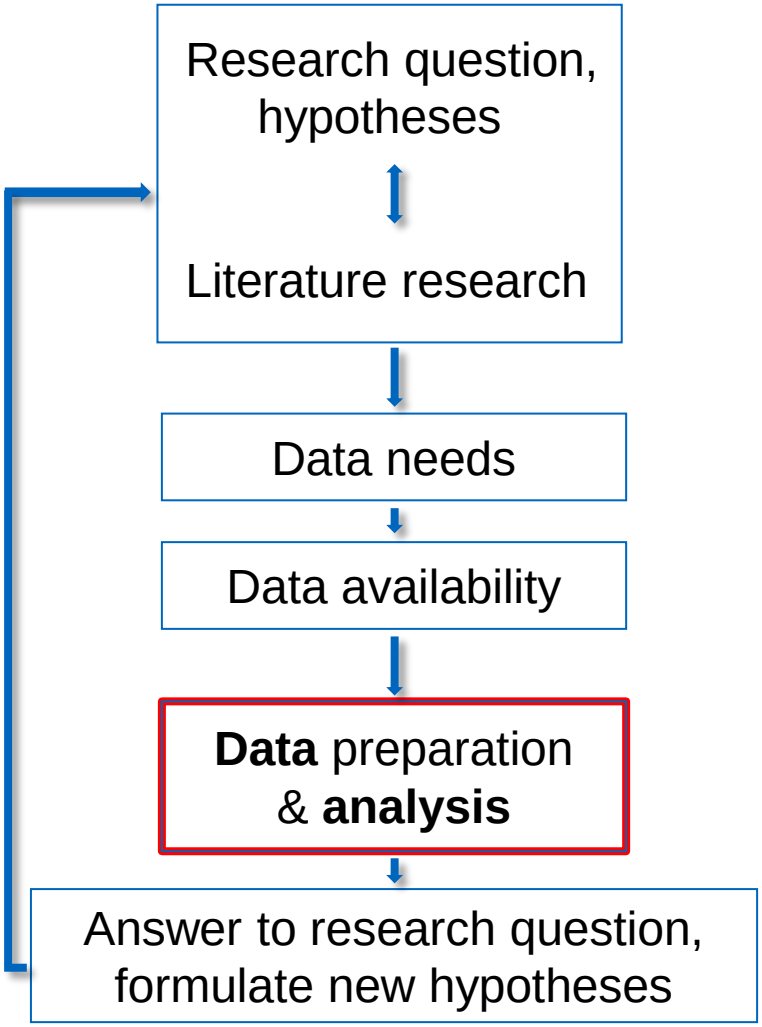
Lecture Plan

Step 3:

- **Data** preparation & **analysis**

Learning objectives

- What is descriptive and inferential analysis?
- Which indicators describe empirical data?



Inferential Statistics

- Descriptive statistics provide information about our immediate group of people (sample), not about all people (population)
- Often you do not have access to population but only to a limited number of people (the sample) instead
- Inferential statistics are techniques that allow us to use these samples to make generalizations about the populations from which the samples were drawn.

Inferential Statistics

There is a wide range of statistical tests available. You need to choose the one that fits with your research question and your data characteristics to get reliable and valid results.

For the choice three questions are important:

1. What are you testing for?
2. What type of data do you have?
3. Do you have related or unrelated data?

Inferential Statistics

1. What are you testing for?
   - Relations vs. Differences
   - One-tailed vs. Two-tailed

2. What type of data do you have?
   - Number of variables
   - Level of scale of variables
   - Distribution of variables (most important normal distribution)

3. Do you have related or unrelated data?
   - Paired vs. independent observations

CASE STUDY

RQ: How does accessibility relate to individual travel behaviour?

- H1: The more often people use the bicycle the better the accessibility by bicycle is assessed for all trip purposes.

H 1: The more often people use the bicycle the better the accessibility by bicycle is assessed for all trip purposes.

1. What are you testing for?
   - Relation between frequency of bicycle usage and assessment of accessibility by bike
   - One-tailed: the more usage the better the assessment

2. What type of data do you have?
   - 1 IV: p033: Frequency of bicycle use
   - I DV: p0412 Accessibility per bicycle working shopping
   - IV ordinal, DV metric
   - DV not normal distributed

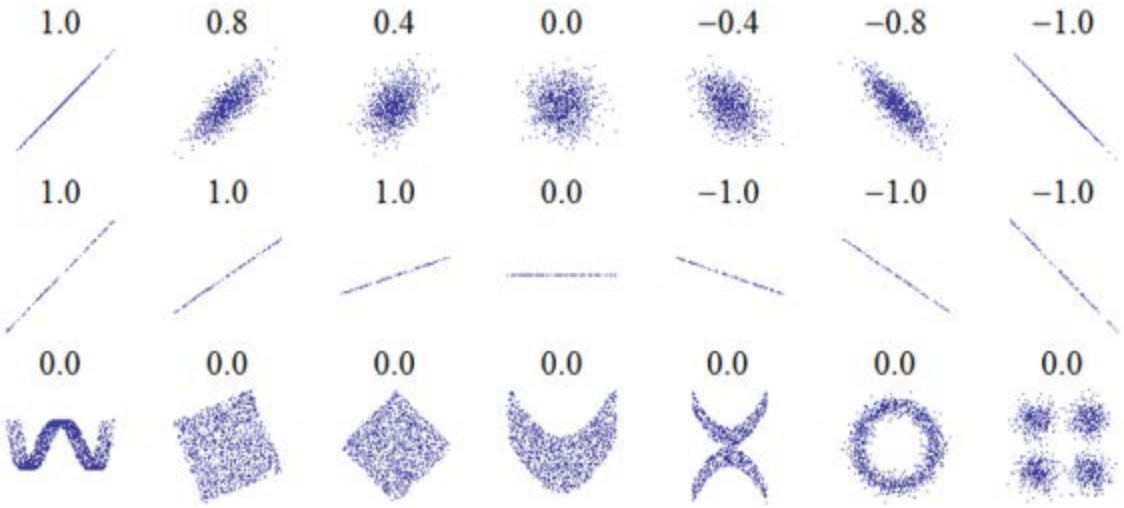
3. Do you have related or unrelated data?
   - IV and DV independent observations

Correlation analysis

- Correlation analysis
- Correlation is a measure of linear relationship
- The Correlation Coefficient is a dimensionless measure of the interdependence between two variables.
- Correlation coefficients state:
  - Whether there is a relationship
  - The direction of the relationship
  - The strengths of the relationship

Properties of Correlation Coefficients:

- $r \in [-1,1]$
- r=0 No linear relationship
- r=1 Perfect positive dependence
- r=-1 perfect negative dependence
- Lacking causality



Types of correlation coefficients

Variable Y\X	Quantitiative X	Ordinal X	Nominal X
Quantitative Y	**Pearson r**	Biserial r_b	Point Biserial r_{pb}
Ordinal Y		**Spearman rho/ Kendalls Tau**	Rank Biserial r_{rb}
Nominal Y		Rank Bisereal r_{rb}	**Phi, L, C, Lambda**

- Spearman more robust than Pearson
- Kendall τ is recommended for small sample sizes, non normally distributed data, unequal scales
- Kendall τ and Spearman are highly correlated, show in most cases the same direction and intensity of relationship

Source: adapted from Bortz & Schuster (2010). Statistik. Springer: Berlin

Case study

- Correlation analysis
- Results:
  - Spearman Rho = .27, p<.01
  - Kendalls Tau = .22, p<.01

- Highly significant correlations at p<.01 marked with **
- Significant correlations with at p.<.05 marked with *

Interpretation of Correlation Coefficients

Interpretation	Cohen (1988)	Brosius (1998)	Bühl / Zöfel (2005)
Very weak		<.20	<.20
Weak	< .10	.40	.50
Medium	.30	.60	. 70
High	.50	.80	.90
Very high		>.80	>.90

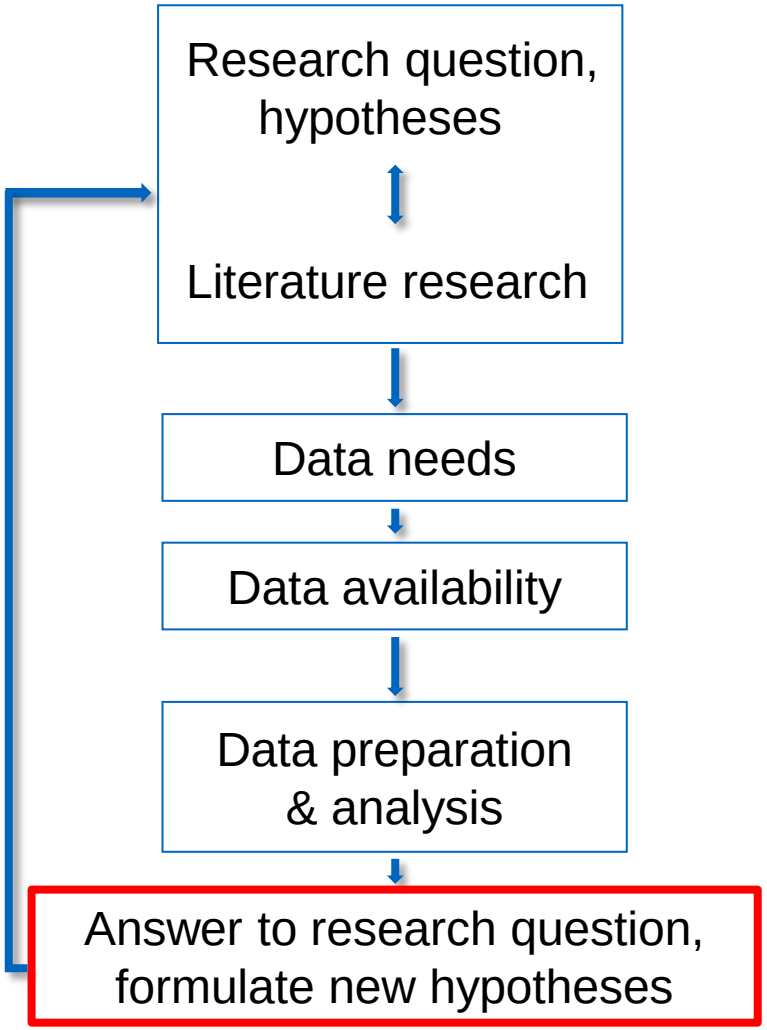
Cohen, J. (1988): Statistical Power Analysis for the Behavioral Sciences, 2. Aufl., Hillsdale: Lawrence Erlbaum Associates
Bühl, A. & Zöfel, P. (2005). SPSS 12. Einführung in die moderne Datenanalyse unter Windows. Pearson: München
Brosius, F. (1998). SPSS 8: International Thomson Publishing

Lecture Plan

Step 4:

- Answer to research question, formulate new hypotheses

- Learning objectives
  - Which information should be provided?
  - How do deal with limitations of the data analysis?



Discussion of results

- Answer the research question based on the results (descriptive & inferential)

- Consider alternative explanations for unexpected results with regards to the:
  - Content
  - Methodology / Statistical analysis

- Develop new hyptheses / research questions

CASE STUDY

RQ: How does accessibility relate to individual travel behaviour?

- H1: The more often people use the bicycle the better the accessibility by bicycle is assessed for all trip purposes.

Descriptive statistics (summary)

H 1: The more often people use the bicycle the better the accessibility by bicycle is assessed for all trip purposes.

Variables	Central tendency	Dispersion	Distribution
Gender	2 = women	Min = 1, Max = 2	47% men 53% women
Age	49.09 years	16.93 (SDDV)	No NV
Frequency of bicycle use	3 = 1-3 days per month	Min = 1 Max = 5	22.0% daily 23.2 % 1-3 d/w, 16.0% 1-3 d/m, 11.3% > m 27.5% never
Accessibility by bike for all modes	2.44 = between 'good' and 'satisfactory'	1.29 (SDDV)	No NV

Inferential statistics (summary)

- Correlation analysis
- Results:
  - Spearman Rho = .27**, p<.01

This corresponds to a positive, weak, statistically significant relation between frequency of bicycle use and assessment of accessibility by bicycle.

Discussion of hypothesis (H1)

- yes, the more often people use the bicycle the better the accessibility by bicycle is assessed for all trip purposes.
- But: the relation is weak
- Possible reasons:
  - Relationship is different for different trip purposes
  - no linear relationship as assumed

Discussion of research question

- Results indicate a weak but positive relation between accessibility and individual travel behaviour for bicyclist.

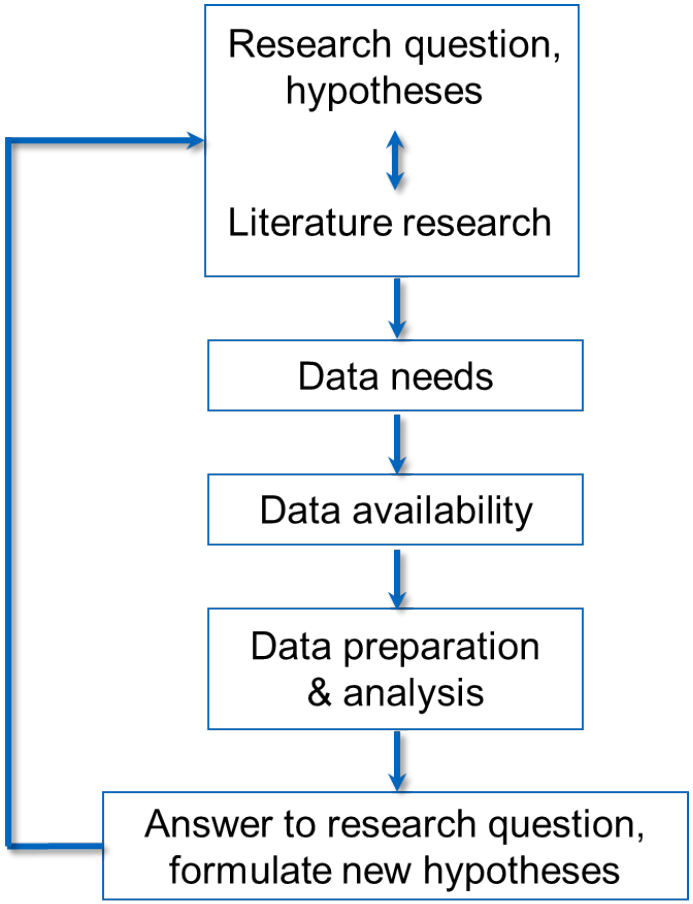
Limitations of the analysis:

- What about other travel modes?
- Only H1, what about H2 & H3

Sum up

Learning objectives

1. Develop an overall understanding of hypothesis-driven data analysis
2. Get to know the relevant steps along the empirical research cycle
3. Be able to conduct a hypothesis-driven data analysis



Even more inferential…Tests

1. What are you testing for?
   - Differences between two groups (means, variance)
   - One-tailed or two-tailed

2. What type of data do you have?
   - 1 IV, 1 DV
   - IV dichotomous (yes/no), DV: numeric
   - DV normally distributed (parametric tests)

3. Do you have related or unrelated data?
   - Paired or independent observations (tests for both cases available)

Tests for a Difference in Means, two Samples

- how likely is it that our two sample means were drawn from the same population?
- If it is highly likely we say that the two sample means are not significantly different.
- If it is rather unlikely we say that the two sample means are significantly different.
- As before we use probabilities to decide what is called likely/unlikely.

- Two important tests for comparing two sample means:
- Student's t-test for independent/paired samples and normally distributed data/large sample sizes.
- Wilcoxon's rank-sum test when the samples are independent/paired but the data is not normally distributed.

Example

- H 3: Car availability relates to the number of trips, trip distance and trip time

Refined for difference testing:

- H3 R: Is there a difference in number of trips (trip distance, trip time) for people with and without a driving license?

Tests for a Difference in Means, two Samples

1. What are you testing for?
   - Differences between people with / without a driving license
   - two-tailed

2. What type of data do you have?
   - 1 IV hp_pkwfs: Car drivers license (yes/no),
   - 1 DV: NoTripsPerson
   - DV normally distributed (to be tested)
   - Equal variances between groups (Homoscedasticity) (to be tested)

3. Do you have related or unrelated data?
   - independent observations