Introduction to quantitative analysis

Petr Ocelík

ESS401 Social Science Methodology

6th October 2015

Outline

- Data and data types
- Descriptive vs. inferential statistics
- Introduction to R

Data

- Concept of data usually taken for granted.
- Data is **information that has been collected and recorded**.
- What we understand as data depends on our philosophical position.
- Quantitative research is typically embedded in positivist tradition.

Types of data

- Qualitative vs. quantitative data.
- Typically described as words vs. numbers.

- Blaikie: all primary data start as words.

Data

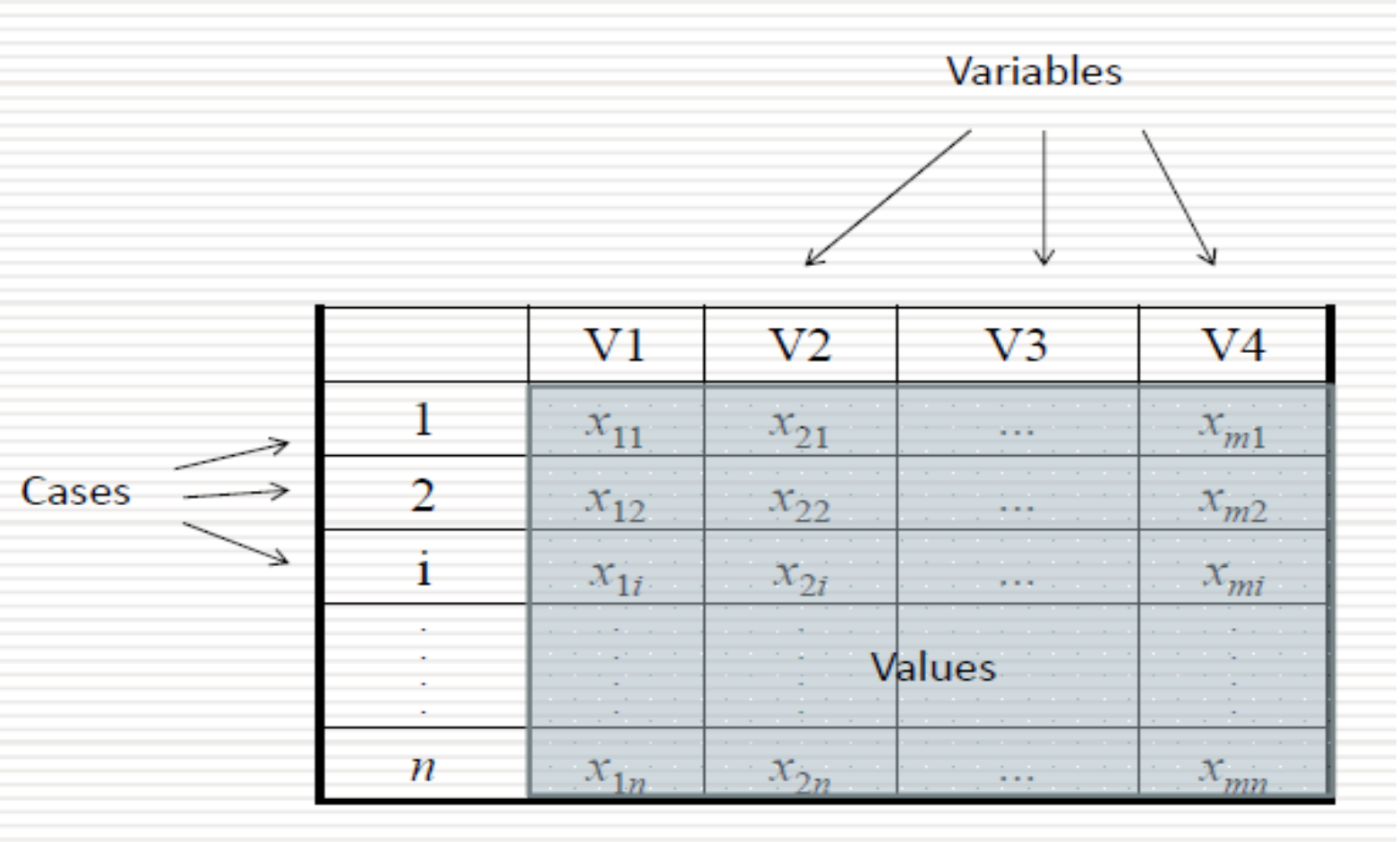
- What does data consist of?

- **Case:** is a unit of observation / analysis.
- **Variable:** is a concept that can have different mutually exclusive values.
- **Value / score:** particular category or point on a measurement scale.

(Kittel 2013)

Everyday usage	Science	Statistics
Object	Unit of analysis, unit of observation	Case
Property	Attribute	Variable
Specific property	Attribute level	Value

Case-by-variable matrix (Kittel 2013)



Discrete and continuous variables

- **Discrete variables:** separate values without intermediate values.
  - Categories (single, married, divorced)
  - Whole numbers (0, 1, 2, …)
- **Continuous variables:** any value over certain interval.
  - Real numbers (0, 1, 1/3, 0.333, 3.14, …)

Levels of measurement

- **Categorical measurement:**
  - Assigns entity to a discrete category.

- **Metric measurement:**
  - Assigns entity to a position at a given numerical scale.

Levels of measurement

- Categorical measurement:
  - Nominal
  - Ordinal
- Metric measurement:
  - Interval
  - Ratio

Nominal measurement

- Construction of categories must be:

- Homogeneous
- Mutually exclusive
- Exhaustive

- We can arbitrarily assign numbers to categories.

Ordinal measurement

- Defined by same conditions as nominal level plus **orders categories along some dimension**.

- We can order categories, but distance between them is not equal.

- Numbers indicate only order of categories.

Interval measurement

- Defined by same conditions as ordinal level plus **scores on a scale are at the same distance apart**.
- We can measure and compare intervals.
- **No true zero:** position of zero is arbitrary.

- Measures how many (counts),
  not how much (ratios).

Ratio measurement

- Defined by same conditions as interval level plus it has **true zero**.

- We can measure and compare ratios / proportions.

- Degrees Celsius (interval) vs. degrees Kelvin (ratio).

(Kittel 2013)

Measurement level	Comparison of characteristics	Comparison of values	Transformations	Examples
Nominal	same/ different	$a = b$, $a \neq b$	unequivocal	Religious denomination, preferred musical style, nationality
Ordinal	bigger/smaller	$a < b$, $a > b$, $a = b$	monotonous	School grades, soccer league, university ranking
Interval	differences	$a - b = c - d$	linear $x' = ax + b$	Temperature in °C, years, IQ-Scale
Ratio	ratios	$a/b = c/d$	proportional $x' = ax$ (defined origin)	Income, age, duration of marriage

What do we mean by statistics

- Broad sense: collection and analysis of numeric data.

- Narrow sense: "*the mathematics of the collection, organization, and interpretation of numerical data, especially analysis of population characteristics by inference from sampling*." (American Heritage Dictionary).

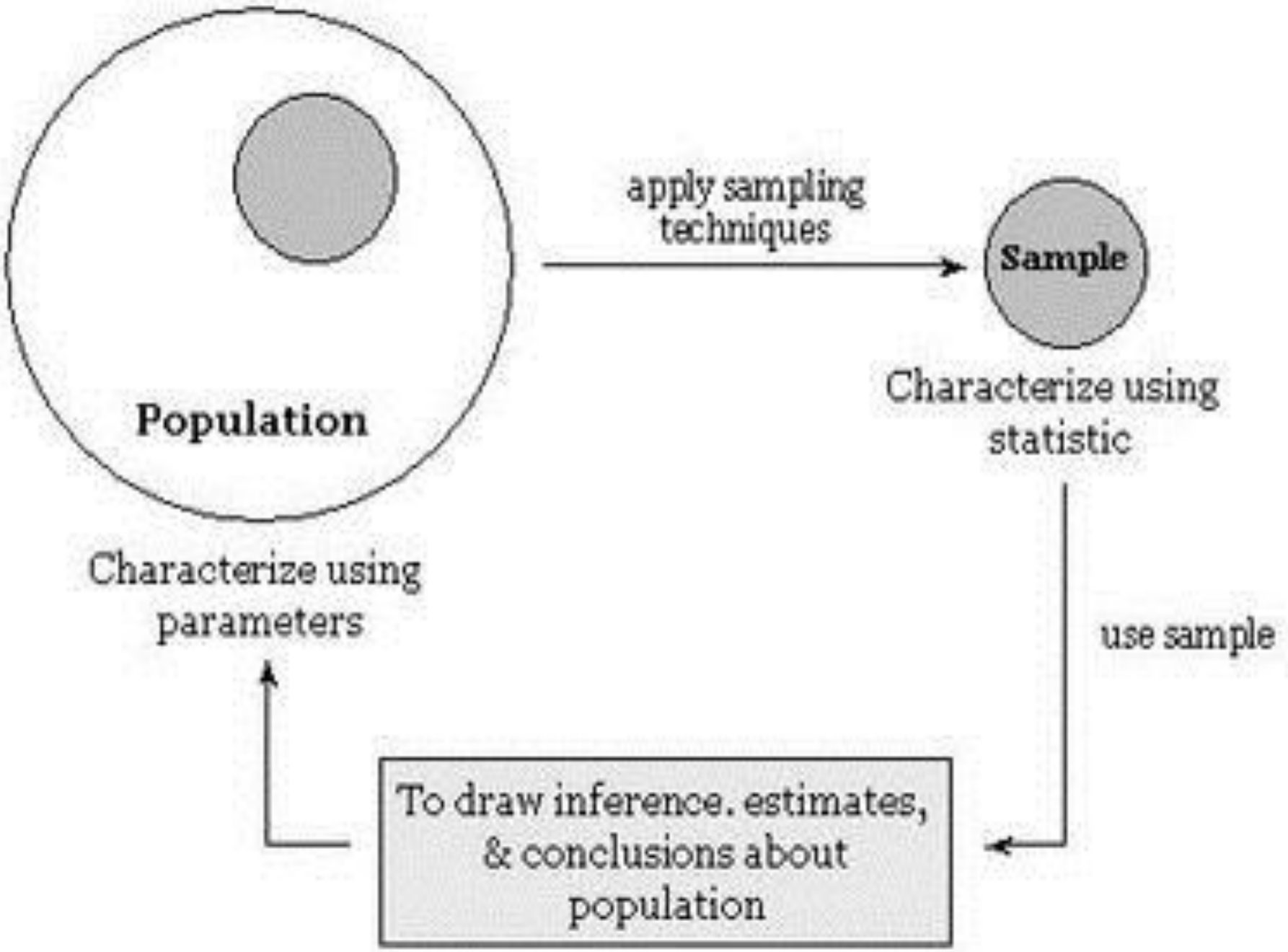
Population and sample

- **Population:** a particular complete set of objects / items with at least one shared property of our interest.
  - **Parameter:** numerical description of population characteristic.
- **Sample:** a subset of population from which data is collected.
  - **(Sample) statistic:** numeric description of sample characteristic.

Descriptive vs. inferential stats

- **Descriptive:** accounts for summarizing and describing data.

- **Inferential:** uses sample statistics to estimate population parameters.


apply sampling techniques
Sample
Characterize using statistic
Population
Characterize using parameters
use sample
To draw inference. estimates, & conclusions about population

Sampling

- **Simple random sample:**
  - Define population
  - Define sampling frame
  - Specify sampling method
  - Choose desired number of items to sample

Sampling variability: R example

- **Population:** sequence of whole numbers from 0 up to 100.
- **Cases:** integers from 0 up to 100.
- Mean of the population (parameter) = 50.
- **Sample:** 10 randomly drawn numbers from the population.
- Mean of the sample (statistic).

Sampling error

- **Sampling error:** the difference between the values of sample statistic and population parameter.
- Sampling error is caused by random selection of the sample.
- There is always sampling error since sample does not include all members of population.

Sampling

- Simple random sample
- Stratified random sample
- Cluster random sample

Non-sampling errors

- **Measurement error:** the difference between measured value of a quantity and its true value.
  - Systematic
  - Random

Non-sampling errors

- **Response bias**
  - Question wording
  - Social acceptability
  - Response set
- **Non-response bias**
  - Refusal to participate in survey
  - Refusal to respond a specific question
  - Missing data

R: advantages

- Freeware
- Open source
- Worldwide active community
- Flexible and developed

R community / sources

- There is huge number of free resources
- R package / library manuals
- R site: http://cran.r-project.org
- Community forums:
  - http://stackoverflow.com
  - http://www.statmethods.net
  - http://www.r-bloggers.com
- Youtube videos: https://www.youtube.com/watch?v=qHfSTRNg6jE
- Googling (often fastest)

R libraries / packages

- Library / package:
  - Can be though of as an extension that adds new functionality.
  - Libraries must be installed (just before the first use) and loaded.
  - Sometimes there can be conflicts among libraries (e.g. different functions with same names) – we can unload them.
  - Often there are dependencies among libraries (some libraries use functions from other libraries).

R: disadvantages

- Not as easily accessible as “clicking-programs”
- Data preparation could be demanding
- Could be slower for large datasets

R language

- object-oriented programming
  - **object:** instance of certain data class that can be manipulated according set of procedures (methods)
- functional-oriented programming
  - **function:** relation that associates input(s) with output(s)

- We can define certain objects and apply functions on them and vice versa.

Data types

- **Numeric:** continuous numeric data (**-1**, **0.5**, **10.49**)
- **Integer:** discrete numeric data (**-1**, **0**, **1**, …)
- **Character:** string values = **“anythingwithinquotes"**
- **Logical:** output of logical operation

  5 > 10 = **FALSE**

  5 < 7 | 7 > 10 = **TRUE**

Data types: factor

- **Factor:** variable that take limited number of discrete values - levels (categorical variable).
- Factor function converts vector of values into vector of **factor values** (always have form of **character**).
- Factors can be **unordered** (nominal variable) or **ordered** (ordinal variable).

```
> data = c(1,2,2,3,1,2,3,3,1,2,3,3,1)
> fdata = factor(data)
> fdata
 [1] 1 2 2 3 1 2 3 3 1 2 3 3 1
Levels: 1 2 3
```

http://www.r-tutor.com/

R: object and function

- **Object:**

  vector <- c(1,2,3,4,5)

- **Function:**

  fun <- function(x) { x^2 }

- **Output:**

  fun(vector) = 1, 4, 9, 16, 25

- **Nesting:**

  fun_2 <- function(x) { fun(x) + 1 }

R functions

- *word()* indicates function
- mean(vector)

- *function(argument_1, argument_2, …)*
- sample(0:100, 10, rep=FALSE)

- basic functions (part of the basic R package)
- package functions (part of the particular package)
- user functions (user-defined functions)

R objects

- **Vector**
  - Sequence (1-dimensional) of elements of **same data type**
- **Matrix**
  - 2-dimensional rectangular collection of elements of **same data type**
  - Array: n-dimensional matrix.
- **List**
  - Vector that can contain elements of **different data types**
- **Data frame**
  - List of vectors of equal length
  - Table data

Vector

```
> c(2, 3, 5)
[1] 2 3 5
```

```
> c("aa", "bb", "cc", "dd", "ee")
[1] "aa" "bb" "cc" "dd" "ee"
```

```
> c(TRUE, FALSE, TRUE, FALSE, FALSE)
[1]  TRUE FALSE  TRUE FALSE FALSE
```

http://www.r-tutor.com/

Matrix

```
     [,1] [,2] [,3]
[1,]    2    4    3
[2,]    1    5    7
```

http://www.r-tutor.com/

List

```
> n = c(2, 3, 5)
> s = c("aa", "bb", "cc", "dd", "ee")
> b = c(TRUE, FALSE, TRUE, FALSE, FALSE)
> x = list(n, s, b, 3)   # x contains copies of n, s, b
```

```
> x[c(2, 4)]
[[1]]
[1] "aa" "bb" "cc" "dd" "ee"

[[2]]
[1] 3
```

http://www.r-tutor.com/

Data frame

```
> mtcars
                   mpg cyl disp  hp drat   wt ...
Mazda RX4         21.0   6  160 110 3.90 2.62 ...
Mazda RX4 Wag     21.0   6  160 110 3.90 2.88 ...
Datsun 710        22.8   4  108  93 3.85 2.32 ...
                  ..............
```

http://www.r-tutor.com/