Capture Recapture

Ana Bierrenbach
WHO /STB / TME
bierrenbacha@who.int

Based on lecture elaborated by Dr. Udo Buchholz

Overview

- Introduction
- How to construct the complete list
- Inventory method
- Principle of capture - recapture
- Assumptions
- 2 sources: formula
- Overestimation, under-estimation
- Tackling assumptions

Introduction

- No surveillance system captures ALL cases
- … but one can try to estimate the real numbers
- There are two methods that one can use if there are two systems that collect data on cases:
  - Inventory method
  - Capture – recapture
- Sources of data:
  - Hospital discharge or admission data
  - Data from vital registration
  - Health insurance data
  - Data from primary care physicians registers within a national health service
  - Drug prescriptions data
  - Data from microscopy services (for AFB+ cases)

Two data sources

- Assume:
  - Health insurance: 100 cases
  - National TB program: 200 cases
- True number?
  - 200?
  - 300? (100 + 200)
  - >300?
- Two ways to tackle the problem:
  - Minimum number: inventory method
  - Estimated number: capture - recapture

Inventory method
step 1: construct the lists

Lista 1: Health insurance

- Ana Bolika, 23 years, female
- Juan Formell, 23 years, male
- Emilio Burrito, 17 years, male
- Mario Maradona, 33 years, male
- Victor Manuel, 49 years, male
- Shaka Shakira, 28 years, female

List 1: 6

List 2: National TB program

- Ana Bolika, 23 years, female
- Feng Shui, 2220 years, female
- Miguel Ballack, 53 years, male
- Mario Maradona, 33 years, male
- Mariana Grajales, 98 years, female
- Hector Sanchez, 44 years, male
- Paolo Almodovar, 45, years, male

List 2: 7

List 3: Complete list

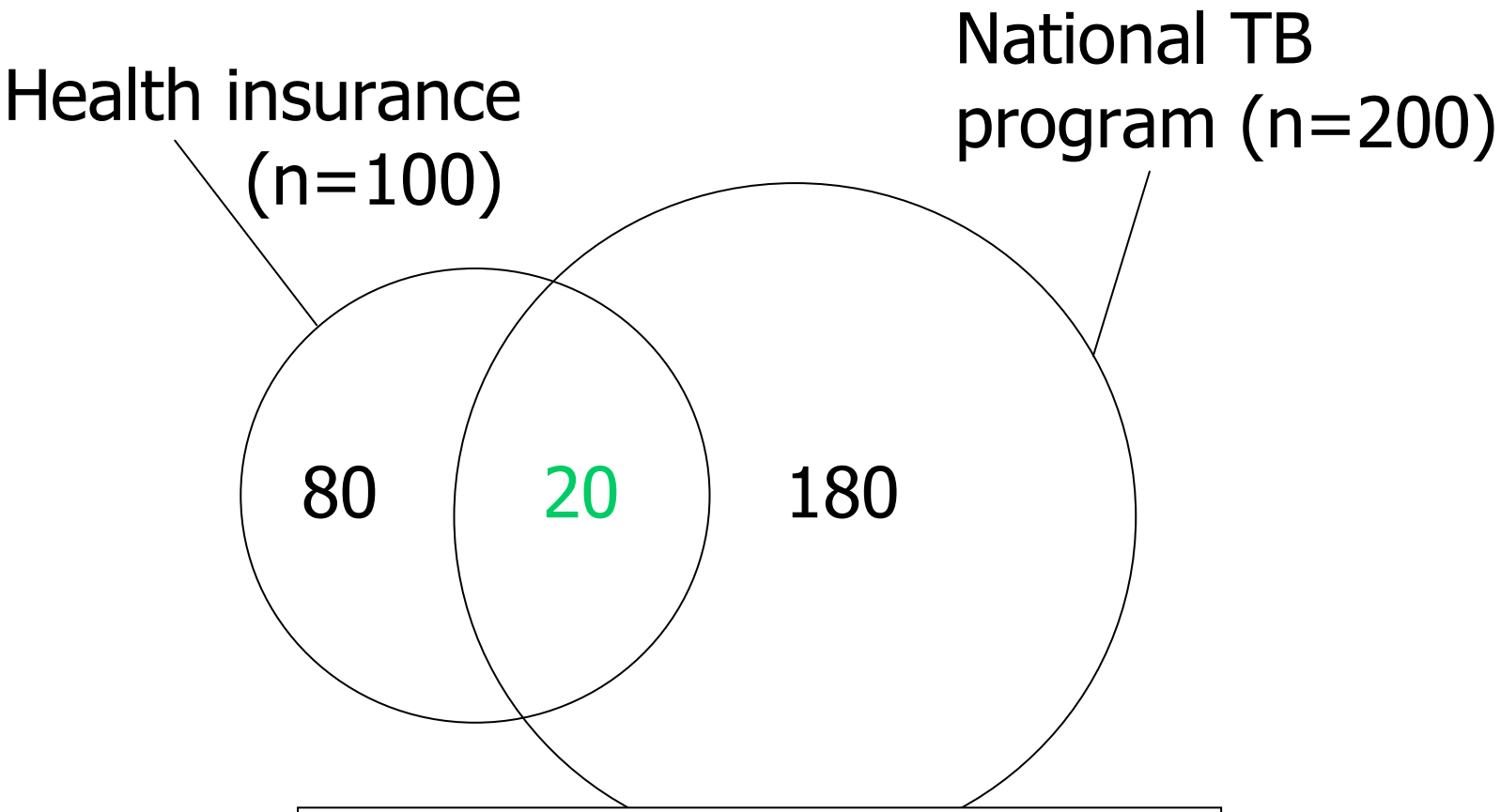
- Ana Bolika, 23 years, female
- Juan Formell, 23 years, male
- Feng Shui, 2220 years, female
- Miguel Ballack, 53 years, male
- Emilio Burrito, 17 years, male
- Mario Maradona, 33 years, male
- Mariana Grajales, 98 years, female
- Victor Manuel, 49 years, male
- Hector Sanchez, 44 years, male
- Shaka Shakira, 28 years, female
- Paolo Almodovar, 45, years, male

List 3: 11

Common list (between 1 & 2): 2

Inventory method
step 2: count the cases

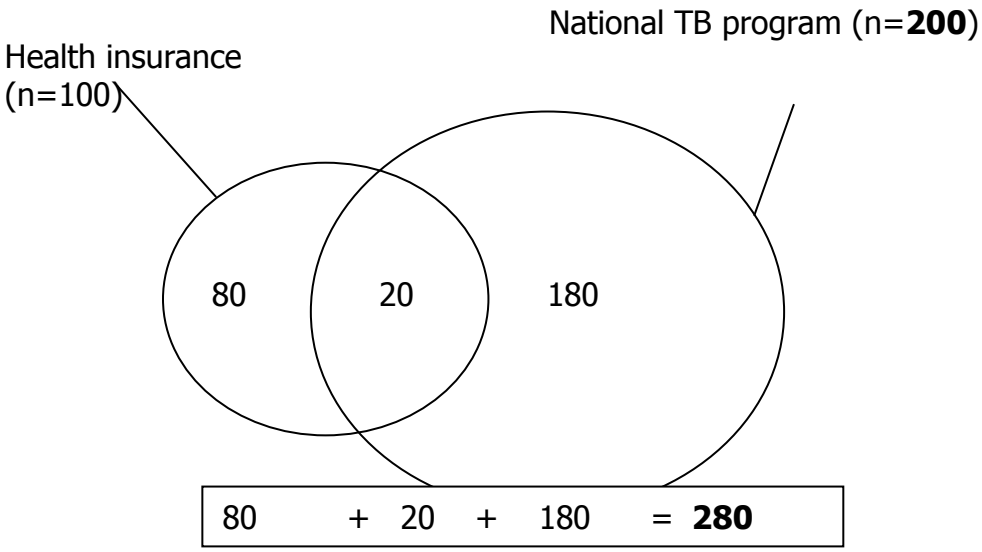
Simply count the cases:



ROBERT KOCH INSTITUT

Calculate multiplyer

„Case detection rate", multiplyer:

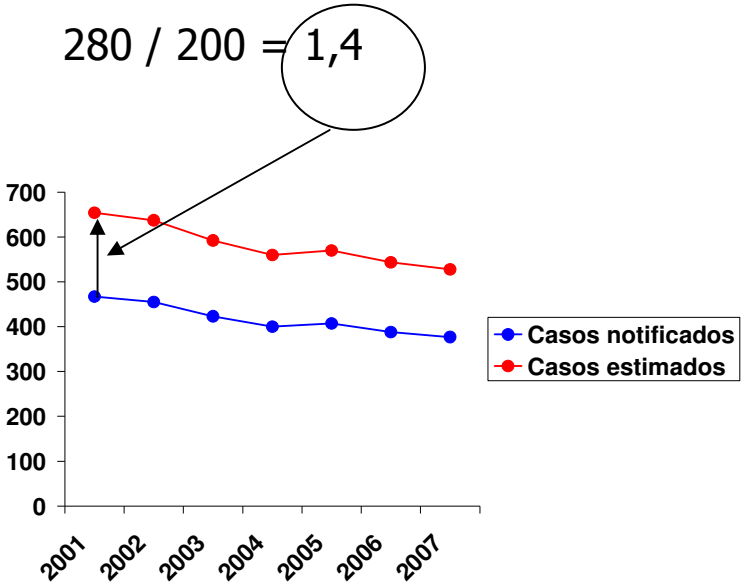


„Case detection rate":

200 / 280 = 71%

Multiplyer:

280 / 200 = 1,4

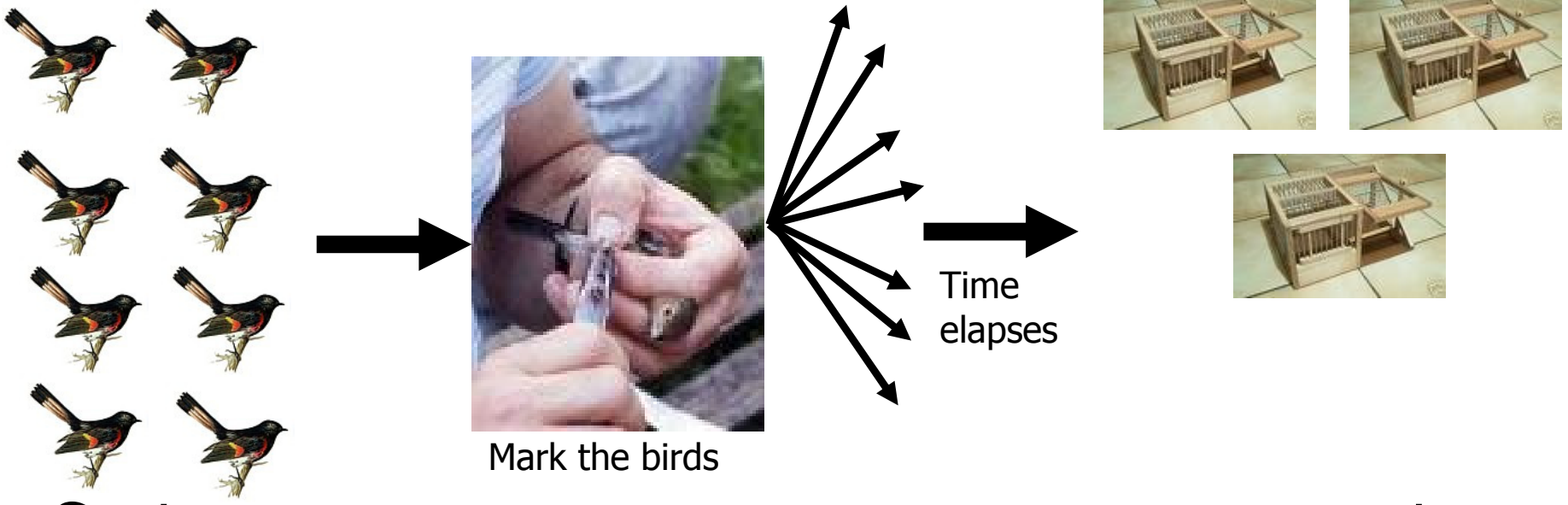


- **May be calculated separately for different age groups**

Principle of capture - recapture

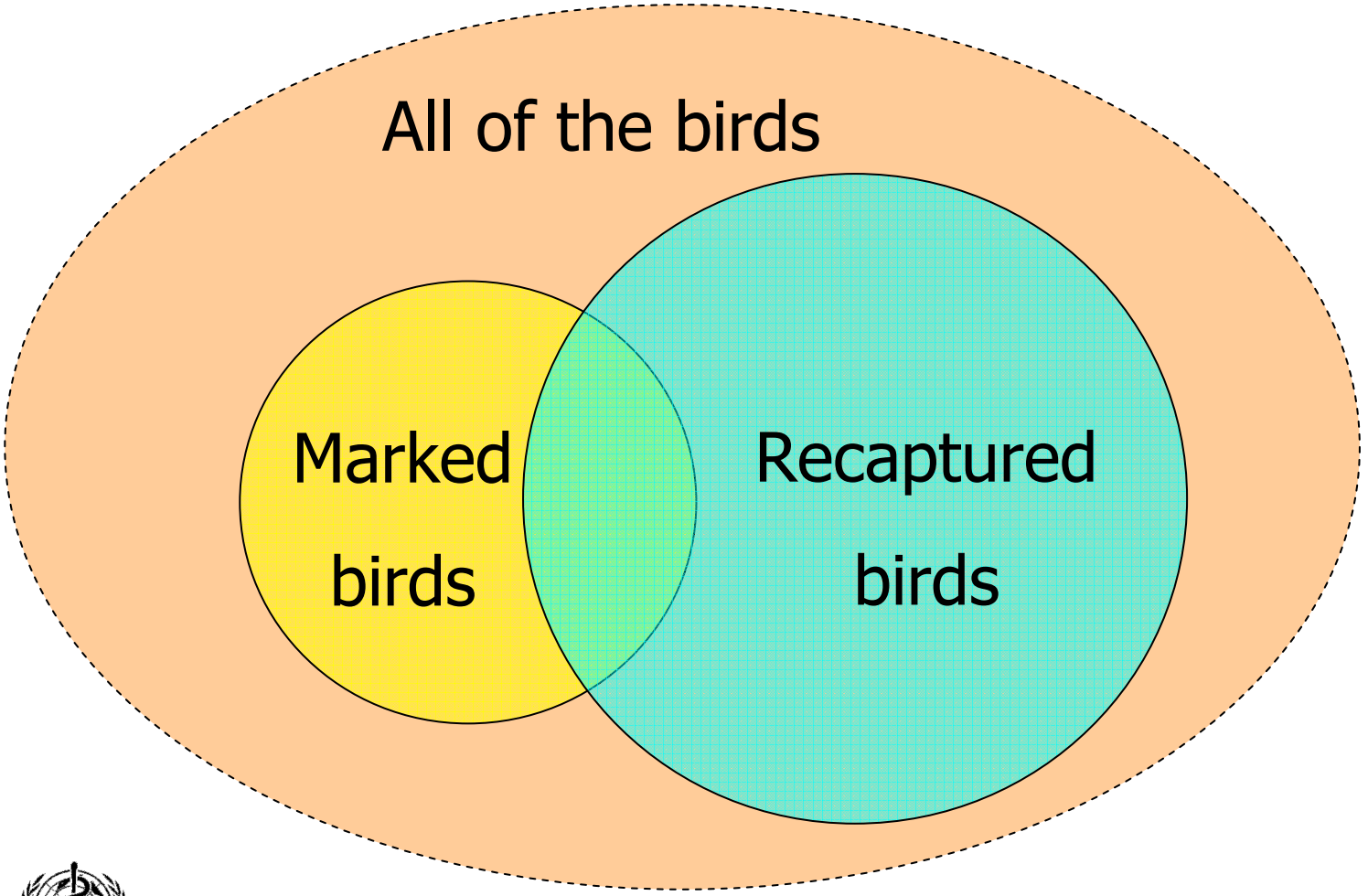
How many birds/mosquitoes… are there in the central park?

- Difficult to count!



ROBERT KOCH INSTITUT

Estimation



2x2 table

	Birds recaptured	- not recaptured	
Marked birds	20	80	100
-not marked	180	??	??
	200	??	??

All of the birds

Assumptions 1

- **Closed population**: no birth, death or migration
- **Perferct linkage**: no matching error
- **No misdiagnosis**: correct case definition
- **Homogeneity capture:** for a given source, every case has the same chance of being captured, the same "catchability"
  - For example: male and female cases, rich and poor cases should have the same chance of being captured

Assumptions 2

- **Source independence:** Visibility vs invisibility to the health care system
- The sources of data have to be independent from each other. That is, cases captured in system 1 should have the same probability to show up in system 2 as those not captured in system 1
  - Negatively dependent:
    - Cases with a health insurance have a 40% chance of being notified to the NTP, but cases without health insurance (e.g. treated in governmental services for the poor) have a 90% chance
    - Mutually exclusive data sources based on geographical criteria
  - Positively dependent: NTP registry and public hospital

2 sources: formula

	B	-	
A	**N_{AB}**	80	**N_A**
-	180	??	??
	N_B	??	**N**

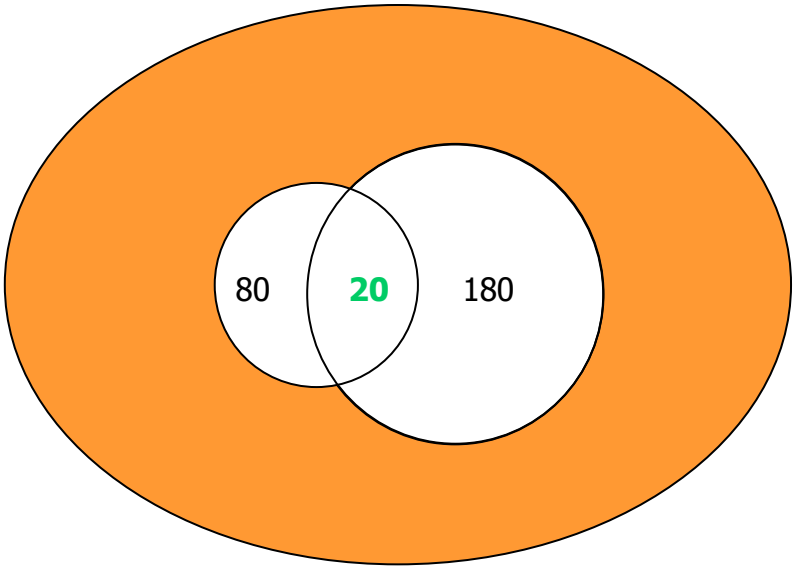
- Only if the two sources are independent: P(A and B) = P(A) x P(B), so that:

$$\frac{N_{AB}}{N} = \frac{N_A}{N} x \frac{N_B}{N} \rightarrow N = \frac{N_A \times N_B}{N_{AB}} \rightarrow N = \frac{(N_A + 1) \times (N_B + 1)}{(N_{AB} + 1)} - 1$$

More exact

Estimation for two sources

	Birds recaptured	- not recaptured	
Birds marked	**20**	80	100
- not marked	180	??	??
	200	??	??



$$N = \frac{(N_A + 1) \times (N_B + 1)}{(N_{AB} + 1)} - 1$$

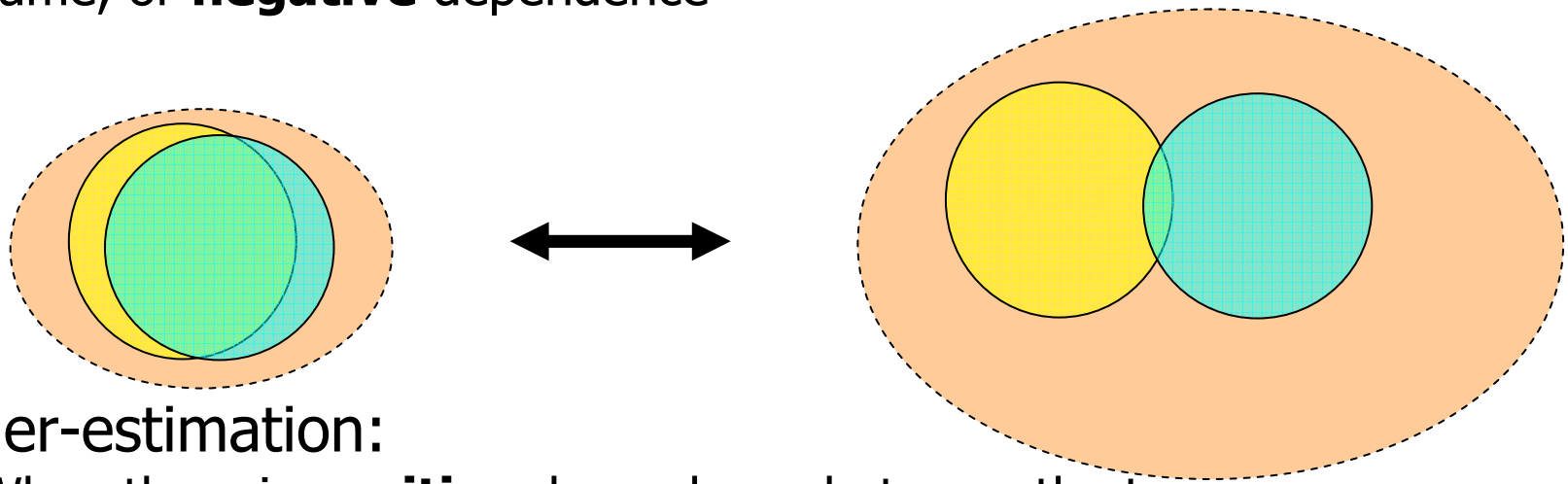
=[(100+1)*(200+1)/(20+1)]-1=**966**

Remember: the inventory method yielded **280** cases

ROBERT KOCH INSTITUT

Overestimation, underestimation

- Over-estimation:
  - When cases captured in both sources are not recognized as being the same, or **negative** dependence



- Under-estimation:
  - When there is **positive** dependence between the two sources
  - Note: it may still be worth doing a capture-recapture study as the estimated number is likely still closer to the truth than just relying on the national notification system. In addition: one always learns a lot about loopholes in the system

Example for overestimation: the difference in the estimates can be great if matching cases are missed

20 matches

	Pajaros recapturados	- no recapturados	
Pajaros marcados	**20**	80	100
- no marcados	180	??	??
	200	??	??

80 matches

	Pajaros recapturados	- no recapturados	
Pajaros marcados	**80**	20	100
- no marcados	120	??	??
	200	??	??

$=[(100+1)*(200+1)/(\mathbf{20}+1)]-1=\mathbf{966}$ $=[(100+1)*(200+1)/(\mathbf{80}+1)]-1=\mathbf{250}$

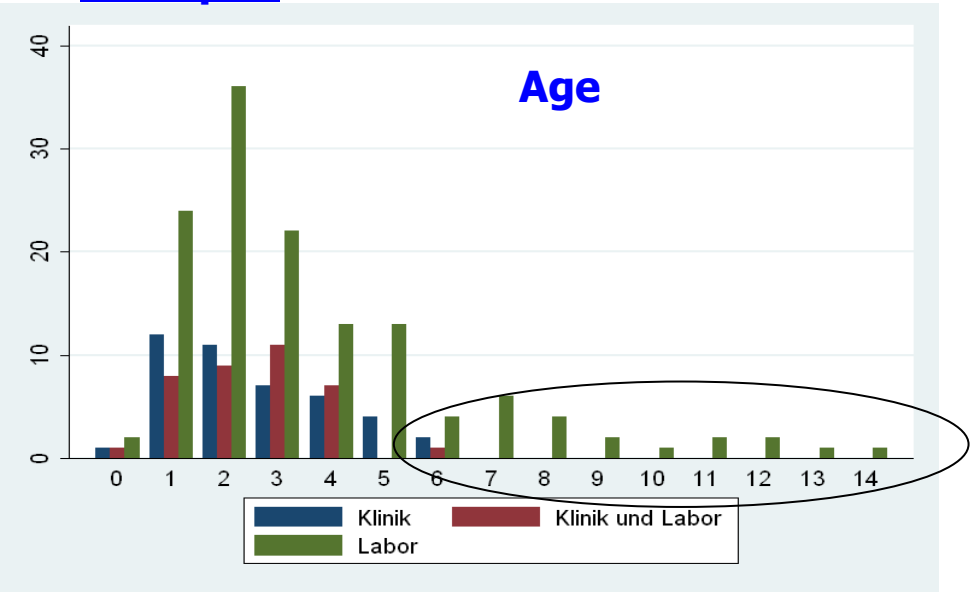
What a difference!!

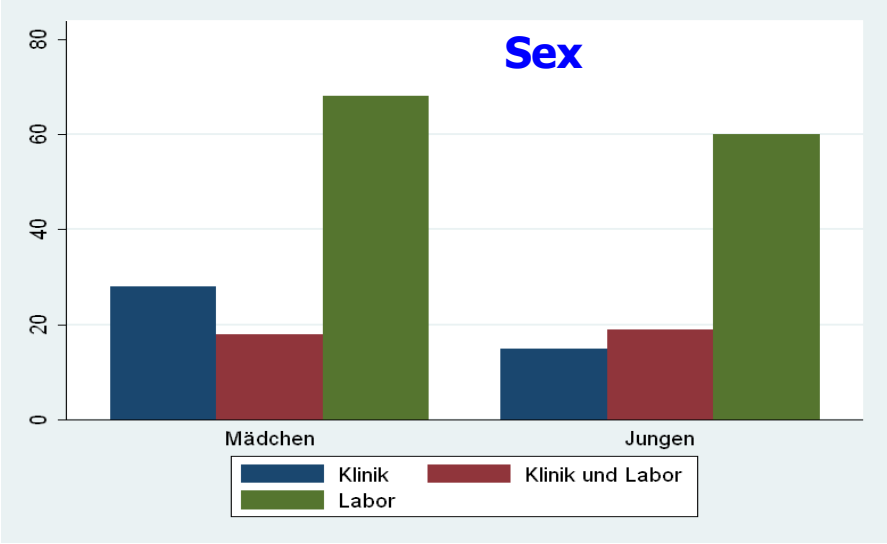
The inventory method would have yielded: **280 cases**

Tackling dependence (1)

- Stratification helps but does not solve the problem

Example:





- It seems that the blue source systematically "loses" children older than 5 years
- Regarding gender, there is no large difference visible

Tackling dependence (2)

- **Modelling: several models**
  - **When there are only two souces one cannot check this assumption in a proper way (other than qualitatively)**
  - **Requirement of at least 3 data source to be able to check for dependencies and to matematically account for them while computing results**
  - **Can be resource intensive (matching, stats, etc)**

Thank you

ROBERT KOCH INSTITUT