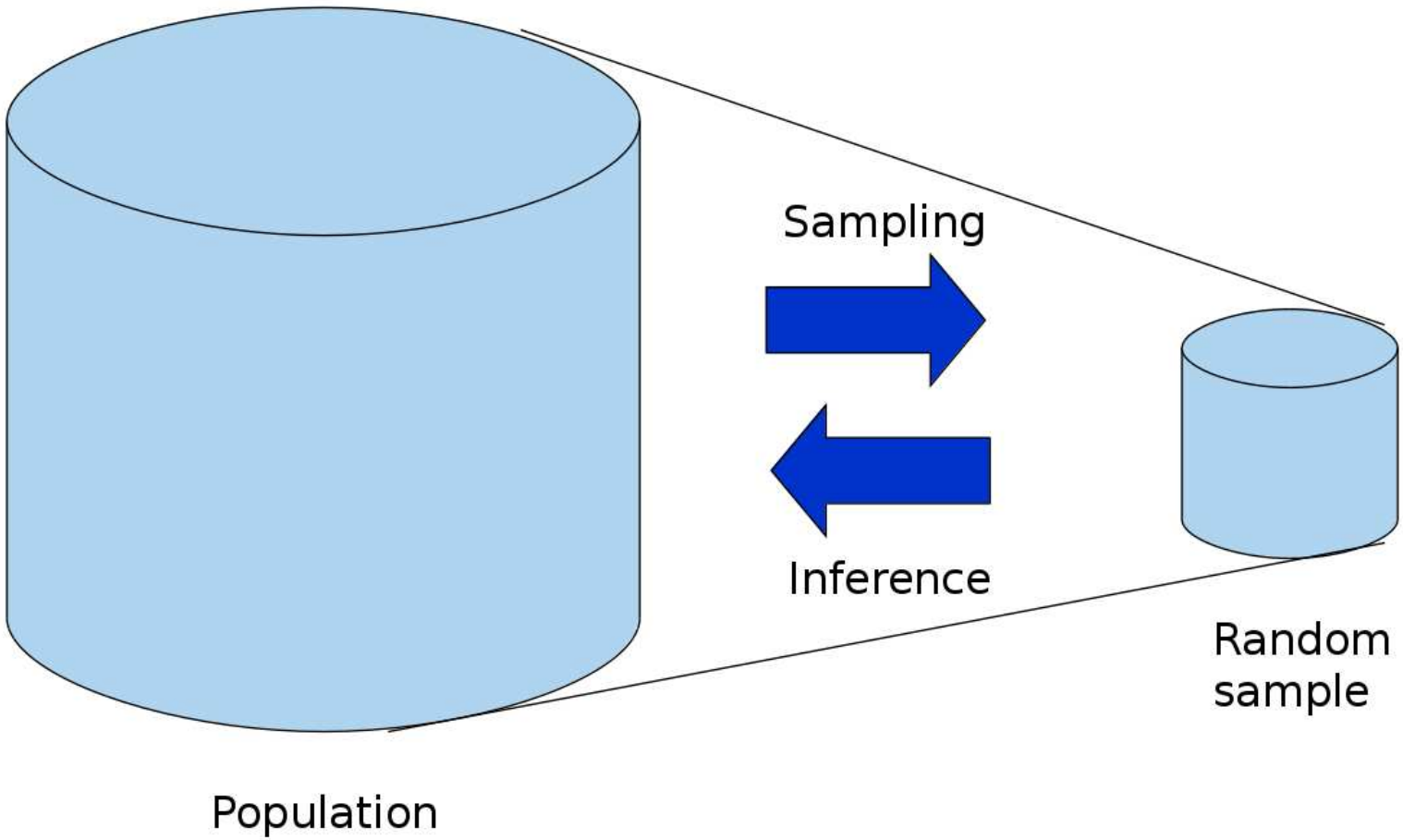
Missing data may bias your conclusions

Juha Karvanen
Department of Mathematics and Statistics
University of Jyväskylä

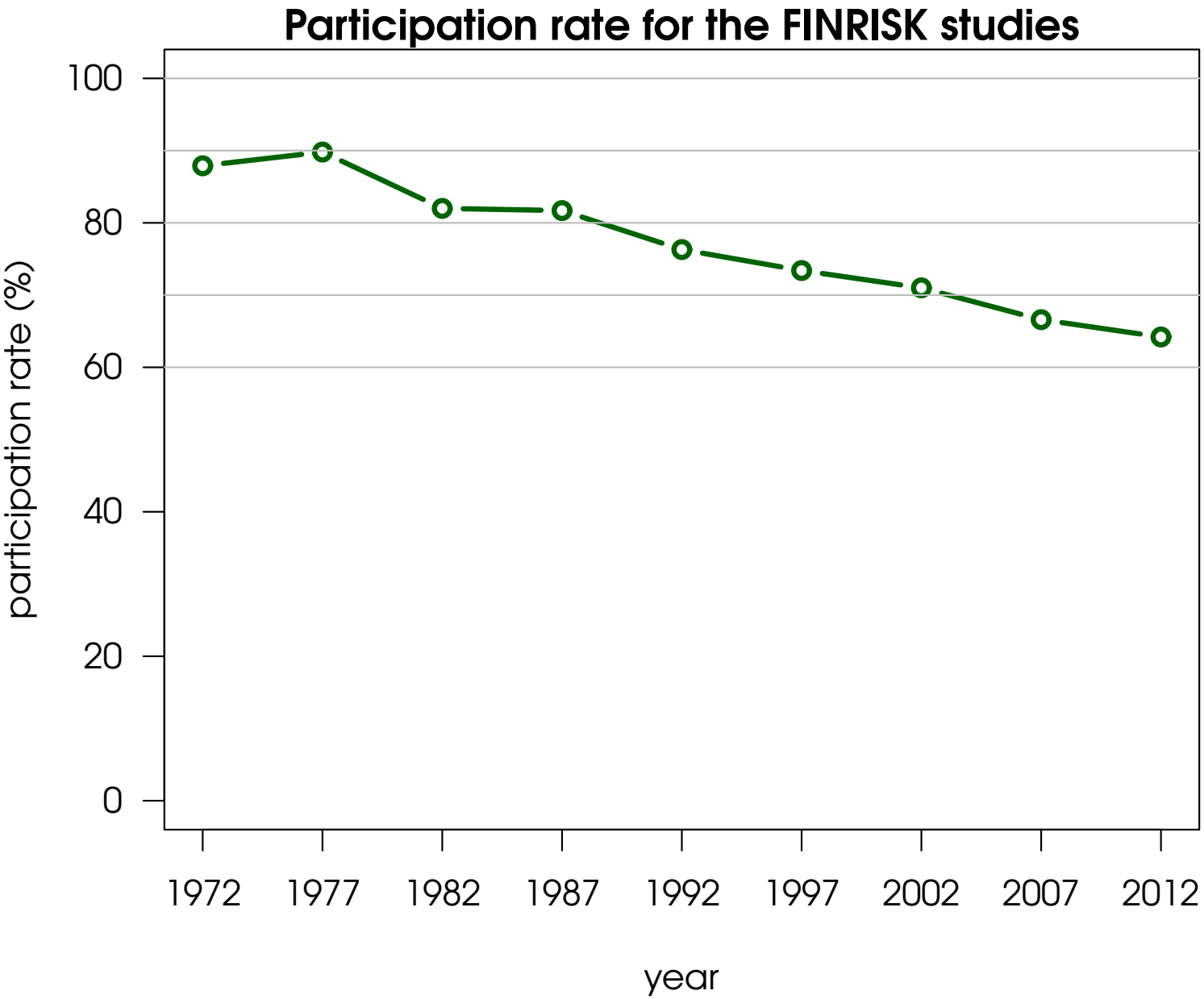
UNIVERSITY OF JYVÄSKYLÄ

Surveys work because of random sampling

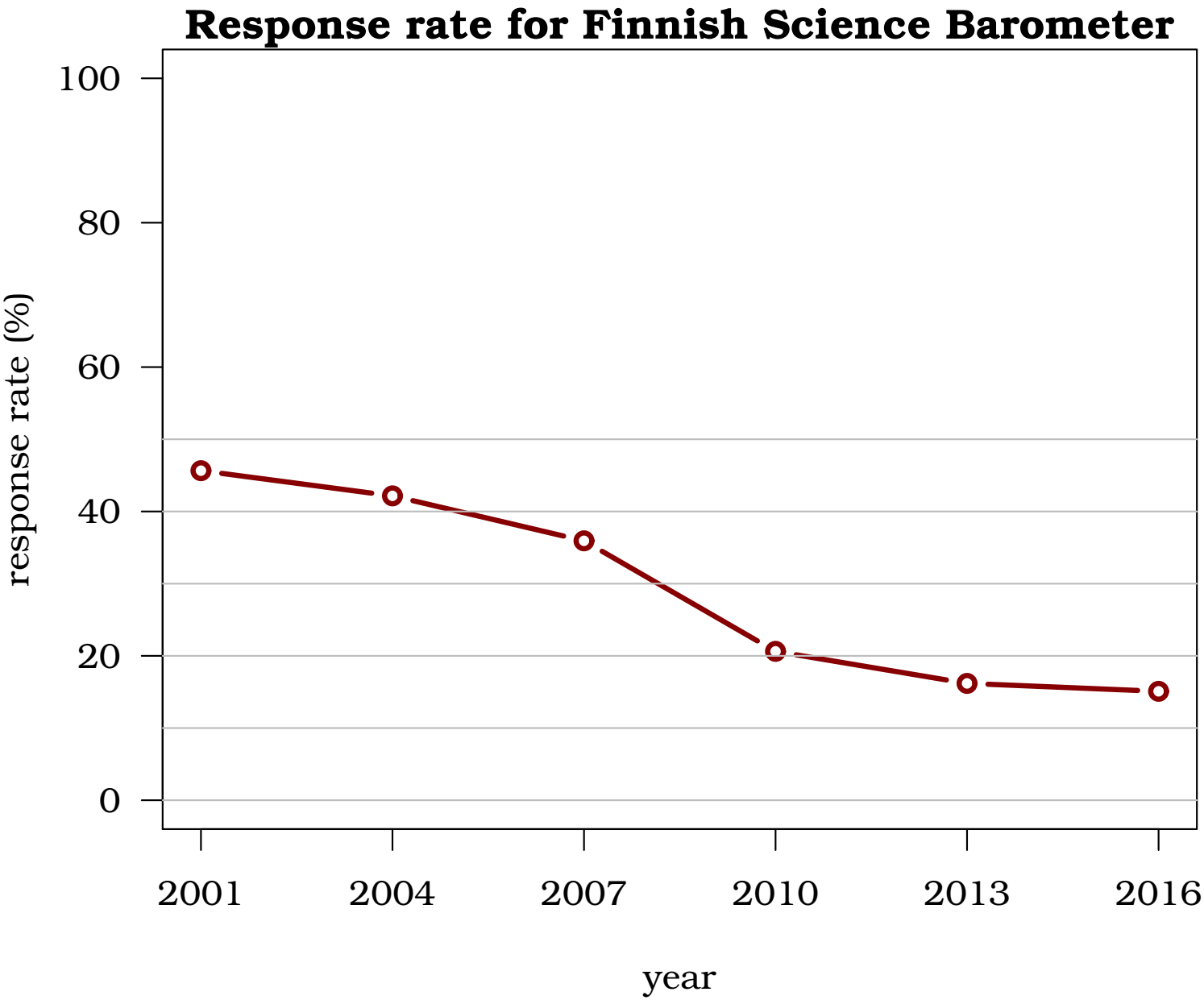


Statistical theory works under the assumption that the data are a random sample from the population. This assumption may be invalid if the data are incomplete.

Response rates have a decreasing trend



The situation may be even worse in postal surveys



Missing data: who cares? (Acatiimi 1/2017 & 2/2017)

Karvanen: Tiedebarometri ei ole luotettava

- ”. . .henkilön asenteet tiedettä kohtaan voivat suoraan vaikuttaa vastaushalukkuuteen.”
- ”Koska vastausosuus on alhainen ja tiedemyönteisten yliedustusta vastaajien joukossa voi pitää vähintäänkin mahdollisena, tiedebarometrin antama kuva suomalaisten tiedemyönteisyydestä ei ole luotettava.”

Kiljunen: Tiedebarometri on luotettava

- ”. . .aineistonkeruussa ei ole edes tavoiteltu – mielipidemittausten valtavirran mukaisesti – maksimaalista vastausprosenttia vaan vastauksia.”
- ”Omien havaintojeni perusteella passiivisuus paikantuu ennen muuta tieteen arvon ja merkityksen itsestäänselvyydeksi ymmärtävään väestön valtaenemmistöön.”
- ”. . .yhtä tiedemyönteisiä tuloksia on saatu useissa eri maissa ja erilaisilla tiedonkeruu-menetelmillä.”

Laaksonen: Tieteelliset periaatteet Tiedebarometriin

- ”Ainoa hyvä puoli on se, että olen sitä käyttänyt huonona esimerkkinä survey-menetelmiä koskeneessa esitelmässäni.”
- ”. . .karhunnan käyttämättömyys on käsittämätöntä.”
- ”Olisin odottanut Kiljuselta vastauskadon analyysiä. Tähän ei ehkä hänellä ollut aineistoa ellei hän kerännyt brutto-otoksesta yksilötason tietoa, ja analysoinut sitä hyvillä tilastollisilla malleilla.”

Respondents and non-respondents differ: Finrisk 2007

	Participants	Non-participants with recontact response	Non-participants without recontact response
N	6257	473	3270
Women, %	**54.1** (52.6,55.6)	54.7 (49.2,60.2)	**44.0** (41.9,46.1)
Mean age, years	**48.9** (48.6,49.2)	47.3 (46.0,48.5)	**44.6** (44.2,45.1)
Number of deaths 2007–2012	166	34	204
Deaths per 1000 (95% CI)	26.5 (22.5,30.5)	71.9 (48.6,95.2)	62.4 (54.1,70.7)
standardized (95% CI)	**22.1** (18.5,25.7)	55.1 (34.6,75.7)	**49.5** (42.1,57.0)

95% confidence intervals are presented in the parentheses.

J. Karvanen, H. Tolonen, T. Härkänen, P. Jousilahti, K. Kuulasmaa (2016) Selection bias was reduced by recontacting non-participants, Journal of Clinical Epidemiology, Volume 76, pages 209-217.

Consequences of missing data

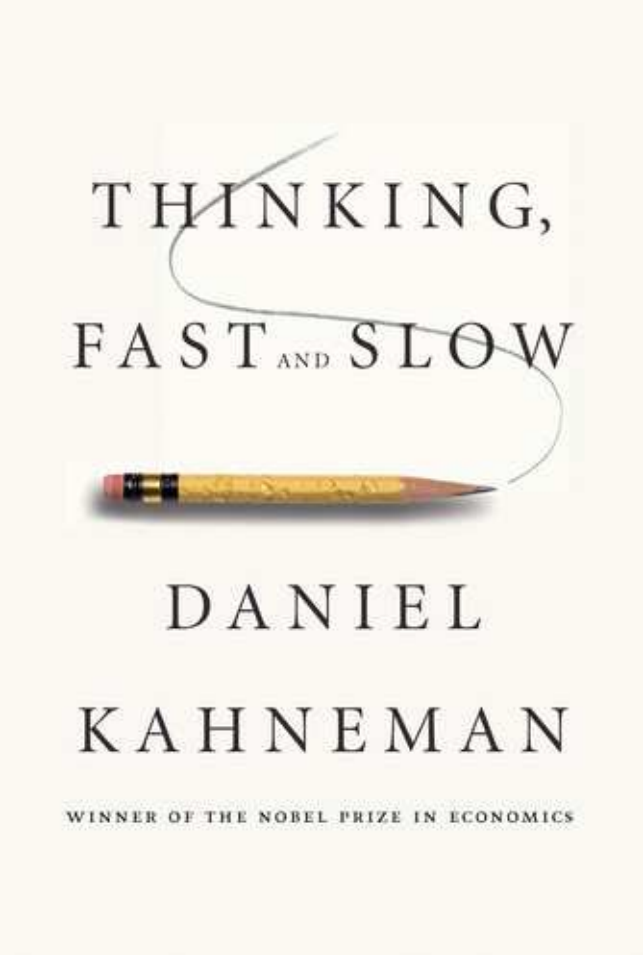
1. Information loss / increased uncertainty
   - The sample size is smaller than anticipated.
   - Confidence intervals are wider.
   - The study may have an insufficient power.
2. Bias
   - Respondents and non-respondents may differ.
   - The sample does not represent the population.
   - Direct averages may have a bias.
3. Bias in trends
   - Changes in time may be due to changes in the response rate.
   - Trends may look more positive they should if the response rate decreases.

Good scientific practice and missing data

"Falsification (misrepresentation) refers to modifying and presenting original observations deliberately so that the results based on those observations are distorted. The falsification of results refers to the unfounded modification or selection of research results. ***Falsification also refers to the omission of results or information that are essential for the conclusions.****"*

"Havaintojen vääristelyllä (falsification, misrepresentation) tarkoitetaan alkuperäisten havaintojen tarkoituksellista muokkaamista tai esittämistä niin, että havaintoihin perustuva tulos vääristyy. Tulosten vääristelyllä tarkoitetaan tieteellisesti perusteetonta tutkimustulosten muuttamista tai valikointia. ***Vääristelyä on myös johtopäätösten kannalta olennaisten tulosten tai tietojen esittämättä jättäminen.****"*

Thinking fast and slow with missing data



Missing data requires slow thinking but people prefer fast thinking.

- Substitutes for thinking:
  - software
  - tradition
  - rules of thumb
  - authoritive authors
  - fancy methods
- An inconvenient truth: analysis of missing data takes a lot of time and effort.
- Key question: **why are the data missing?**

Types of missing data mechanisms

- MCAR (missing completely at random)
- MAR (missing at random)
- MNAR (missing not at random)

Tribute to Hans Rosling 1948–2017

Methods for handling missing data

- Measure again!
- Complete case analysis
- Weighting
- Multiple imputation
- Modeling (Bayesian or likelihood inference)

Random sampling vs. web data

- A small dataset $n = 100$ is collected from the Finnish population ($N = 5000000$) using simple random sampling. The population mean of variable X is estimated by the sample mean.
- Alternatively, it is possible to calculate the sample mean from a web-based dataset. The web data is not a random sample but a self-selected sample, e.g. a volunteer sample. Assume that the correlation between X and the inclusion probability $p(X)$ is known to be 0.05 (weak correlation).
- You may choose whether to use the random sample or the web data. The quality of the estimate is measured by mean squared error which balances the bias and the variance. What at least should be the size of the web data?

A) 120
B) 500
C) 3000
D) 60 000
E) 1 000 000
F) 5 000 000

Meng, X. L. (2014). A trio of inference problems that could win you a Nobel Prize in statistics (if you help fund it). Past, Present, and Future of Statistical Science, 537-562. Boca Raton, FL: CRC Press.

Summary

1. Recognize missing data as a potential problem.
2. Investigate the pattern of missing data.
3. Consider possible reasons why the data are missing.
4. Include handling of missing data as a part of your data analysis. Ask for help if needed.
5. Report all relevant details: the amount of missing data, the potential reasons, the assumptions on the missing data mechanism, the methods used for the analysis, etc.