# The ever changing landscape of survey data collection

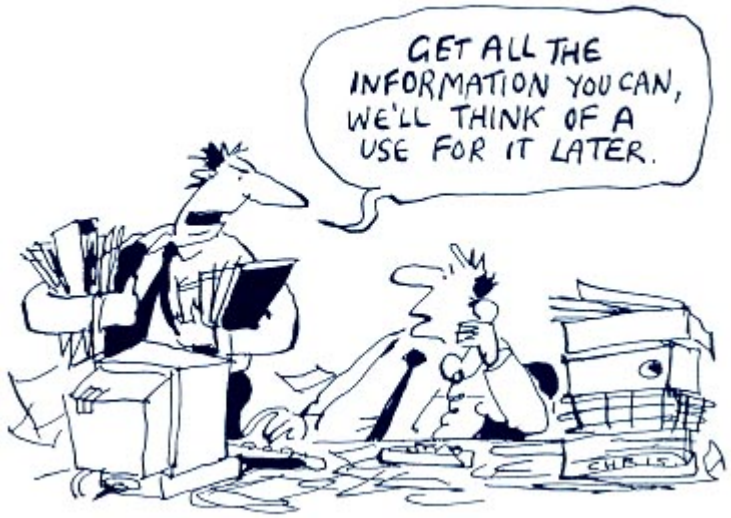
Jelke Bethlehem

Leiden University | Institute of Political Science

# The ever changing landscape of survey data collection

## Overview

- Some history of survey data collection.
- The rise of web surveys.
- The challenges of web surveys.



# Survey data collection through the ages

## Traditional data collection with paper questionnaires

- Face-to-face surveys.
- Telephone surveys.
- Mail surveys.

## Computer-assisted interviewing (1980s)

- Computer-assisted personal interviewing (CAPI).
- Computer-assisted telephone interviewing (CATI).
- Computer-assisted self-interviewing (CASI).

## Internet (1990s)

- Email surveys.
- Web surveys, web panels.

# Sampling for surveys

## The fundamental principles of sampling

- Samples must be selected by means of *probability sampling*.
- Every element must have a *positive* probability of selection.
- All selection probabilities must be *known*.

## Consequences

- It is always possible to construct an *unbiased* (valid) estimator.
- Estimators often have a (approximately) *normal* distribution.
- *Precision* of estimators can be computed (confidence interval, margin of error).

## Warning

- For other forms of sampling (e.g. quota sampling, self-selection, river sampling), it is not clear how valid and precise the outcomes are.

# The challenges of web surveys

## Why are web surveys so attractive?

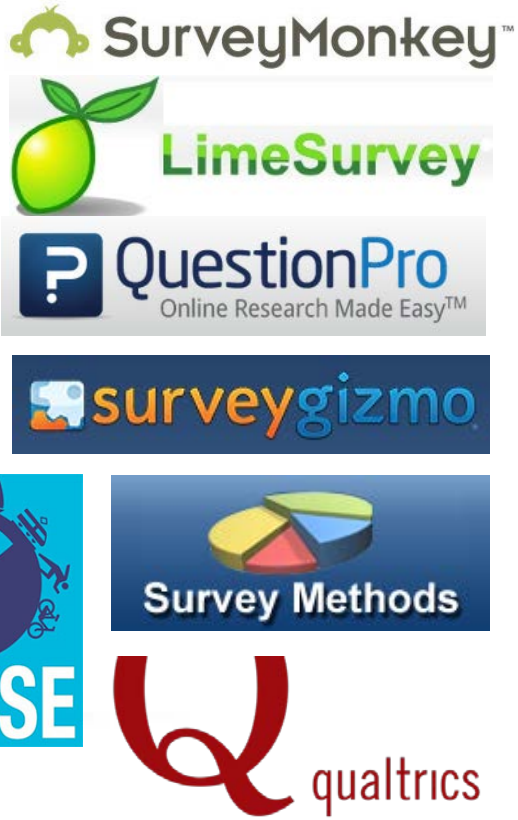
- Easy: simple access to large group of potential respondents.
- Cheap: no interviewers, no printing, no mailing.
- Fast: a poll can be launched very quickly.
- Everybody can do it!

## Methodological challenges

- Under-coverage problems.
- Sample selection problems.
- Nonresponse problems.
- Measurement errors (not discussed).

## The question

- Can web surveys be used in a scientifically sound way?



# Under-coverage problems

## Under-coverage

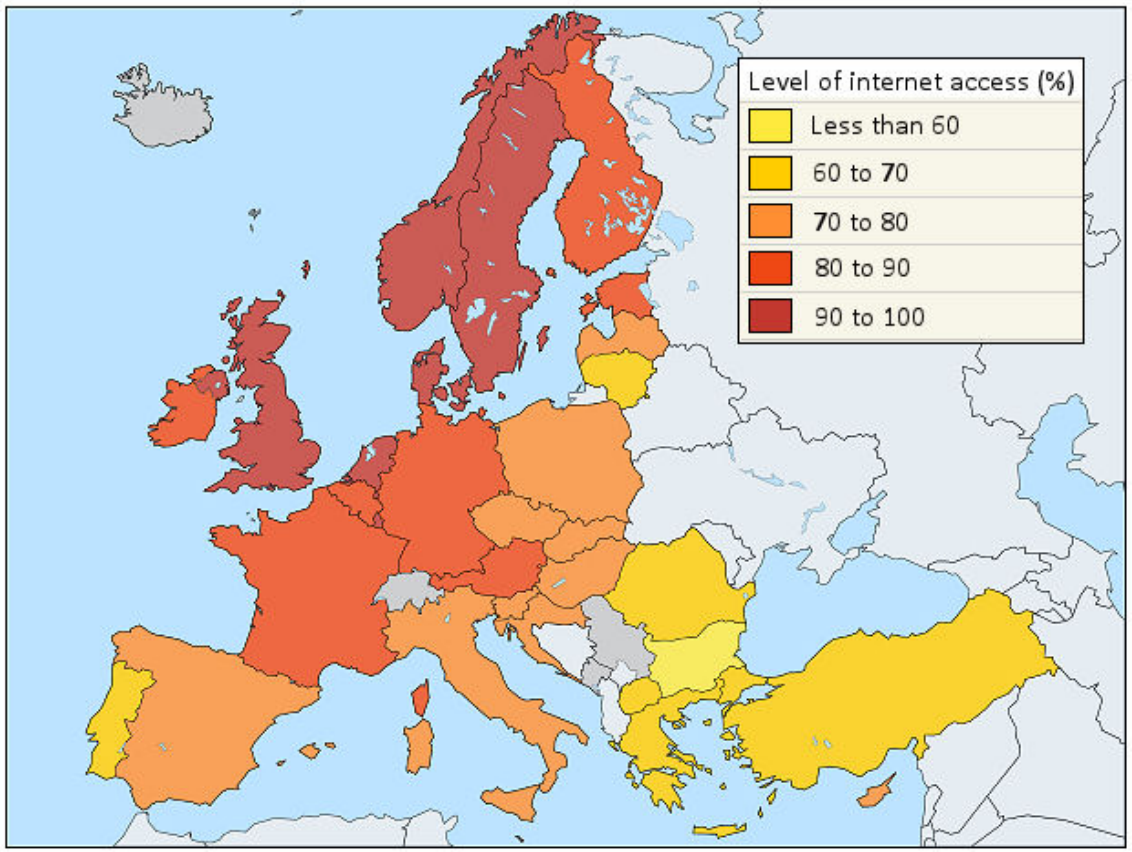
- The target population of a web survey is often much wider than just those having access to the internet.
- Those without internet may differ from those with internet.
- People without internet will never be selected for the survey.
- Therefore, estimates based on web surveys are often biased.

## When is under-coverage a problem?

- For general population surveys.
- Not for, for example, for a survey among students of a university, or employees of a firm. They all have access to internet, and they all have an email address.

# Under-coverage problems

## Internet-coverage in Europe in 2015




**Top 3:**
Norway (97%)
Luxemburg (97%)
Netherlands (96%)

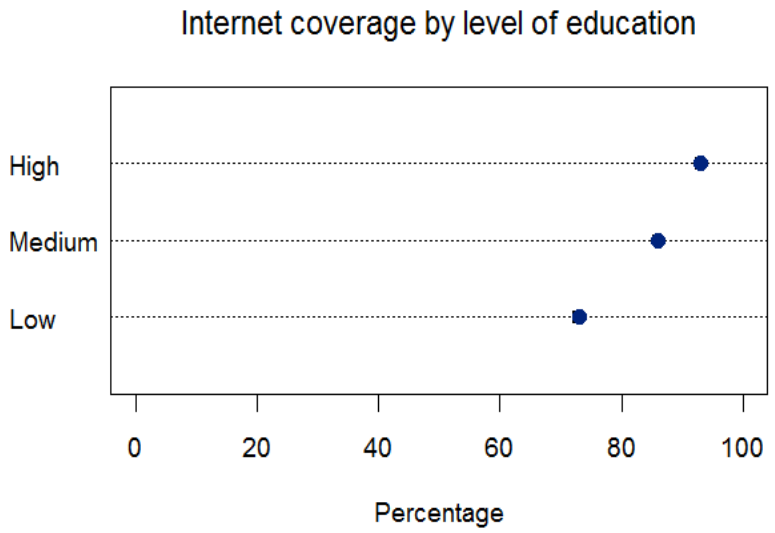
**Bottom 3:**
Greece (68%)
Romania (68%)
Bulgaria(59%)

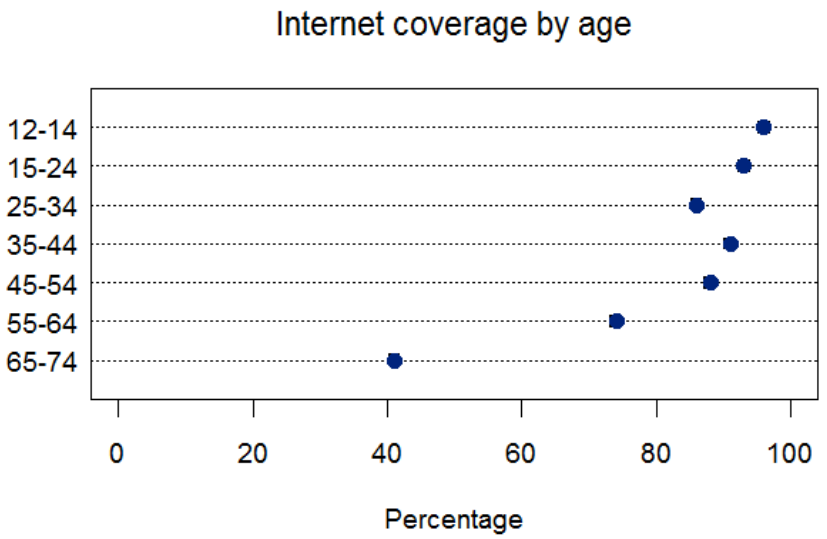
*Source: Eurostat*

# Under-coverage problems

## Groups with lower internet coverage

- The elderly (only 34% for 75+ in the Netherlands in 2013).
- The low-educated.
- Ethnic minority groups.
- Internet coverage by group in the Netherlands (2005):



# Under-coverage problems

## Bias due to under-coverage

$$B(\bar{y}_I) = \frac{N_{NI}}{N}\left(\bar{Y}_I - \bar{Y}_{NI}\right)$$

## The size of the bias is determined by

- The  relative size of the group of people without internet.
- The contrast: the average difference between people with and without internet.

## Note

- The bias need not diminish if coverage increases.

# Under-coverage problems

## Possible solutions

- Wait until internet coverage is sufficiently high.
- Mixed-mode survey. Approach those without internet in a different mode. For example, use CAPI for the elderly. Beware of mode effects.
- Provide free internet access to those without it. Examples: LISS Panel (Netherlands) and KnowledgePanel (US).
- Provide all respondents with a tablet. Example: ELIPSS Panel (France). Advantage: all respondents use same data collection device.



# Sample selection problems

## Selection of a random sample for a web survey

- A sampling frame is required for a probability sample.
- Often, there is no sampling frame of e-mail addresses.
- So it is not possible to send an e-mail with a link to the questionnaire website.

## Alternative: different mode for recruitment

- Draw a random sample from a population register, or an address list, and send a letter (with a link) to each selected person/address.
- Draw a random sample of telephone numbers, call the selected people, and give them a link.
- Disadvantages: more cumbersome, not so fast, increased costs.

## Bad alternative

- Rely on *self-selection* (opt-in) of respondents.

# Sample selection problems

## What is self-selection?

- Form of non-probability sampling.
- Participants are people that have internet, happen to see the invitation, and spontaneously decide to participate.
- It is a cheap and fast way to collect a lot of data.
- However, the sample is usually not representative.

## Other problems

- Also people outside the target population of the survey can respond.
- Often people can respond more than once (on the same or on a different computer).
- Groups of people may attempt to manipulate the outcomes of the web survey.

# Self-selection problems

## Example

- Local elections in Amsterdam in 2014, debate between party leaders.
- Online poll: who was the best?
- Two campaign teams discovered one could vote more than once.
- They voted all night. Results:

| Party | Votes |
|---|---|
| D66 | 3,890 |
| SP | 3,816 |
| PvdA | 1,121 |
| GL | 852 |
| VVD | 214 |

- The poll was cancelled.

# Self-selection problems

## Self-selection bias

$$B(\bar{y}_S) = \frac{R_{\pi Y} S_\pi S_Y}{\bar{\pi}}$$

## Self-selection bias is determined by

- The magnitude of the participation probabilities $\pi_1, \pi_2, \ldots, \pi_N$. The smaller the average participation probability $\bar{\pi}$, the larger the bias.
- The strength of the correlation $R_{\pi Y}$ between participation behaviour and the target variable of the survey. The stronger the correlation, the larger the bias
- The variation (standard deviation) $S_\pi$ of the participation probabilities. The larger the variation, the larger the bias.

# Nonresponse problems

## Nonresponse

- Persons who are selected in the sample (and who belong to the target population) do not provide the requested information.

## Consequences

- Response maybe selective, leading to biased estimates.

## Causes

- *No-contact*: depends on mode of recruitment.
  For example: spam filter.
- *Refusal*: no interest, intrusion of privacy, no time.
- *Not-able*: illness, language problems, no internet.

## Response rates

- Response rates are low in web surveys, often not more than 30%.

# Nonresponse problems

## Nonresponse bias

$$B(\bar{y}_R) = \frac{R_{\rho Y} S_\rho S_Y}{\bar{\rho}}$$

## Nonresponse bias is determined by

- The magnitude of the response probabilities $\rho_1$, $\rho_2$, …, $\rho_N$. The smaller the average response probability $\bar{\rho}$, the larger the bias.
- The strength of the correlation $R_{\rho Y}$ between response behaviour and the target variable of the survey. The stronger the correlation, the larger the bias.
- The variation (standard deviation) $S_\rho$ of the response probabilities. The larger the variation, the larger the bias.

# Self-selection problems and nonresponse problems

## Probability sample + nonresponse

- The maximum absolute bias cannot exceed $B_{max} = S_Y \sqrt{\frac{1}{\bar{\rho}} - 1}$

## Self-selection sample

- The maximum absolute bias cannot exceed $B_{max} = S_Y \sqrt{\frac{1}{\bar{\pi}} - 1}$

## Example

- Statistics Netherlands, probability sample, response rate = 60%: $B_{max} = 0.82 \times S_Y$.
- Self-selection online poll, *21minutes.nl*, $n$=170,000, $N$=12,000,000: $B_{max} = 8.34 \times S_Y$.
- The bias of the online poll can be 10 times as large!

# Reducing the selection bias

## Adjustment weighting

- Assign weights to respondents to correct for over-represented or under-represented groups.
- Weighting techniques: post-stratification, generalized regression estimation, raking ratio estimation, use of propensity scores.

## Required: auxiliary variables

- Must be measured in the survey
- Population distribution must be available.
- They must be correlated with the target variables of the survey.
- They must be correlated with participation behaviour.
- Such variables are often not available.
- So weighting is not always effective.

# Example 1: Shopping Sundays

## Shopping Sundays in the municipality of Alphen a/d Rijn

- Urban town Alphen (70,000 people)
- Seven rural villages: Aarlanderveen, Benthuizen, Boskoop. Hazerswoude-Dorp, Hazerswoude-Rijndijk, Koudekerk, Zwammerdam (together 30,000 people).


*Alphen*

*Benthuizen*

# Example 1: Shopping Sundays

## Shopping Sundays in Alphen a/d Rijn

- Should the shops be open on Sunday?
- Liberal parties in favour, Christian parties opposed.

## Three surveys at the same time

- *Face-to-face interviews* by members of political parties in shopping centres on one Saturday afternoon. 754 interviews.
- *Citizen panel*, based on random sample from population register. 857 interviews. Response: 54%.
- *Self-selection web survey*, to give everyone the possibility to express an opinion. 1550 interviews.

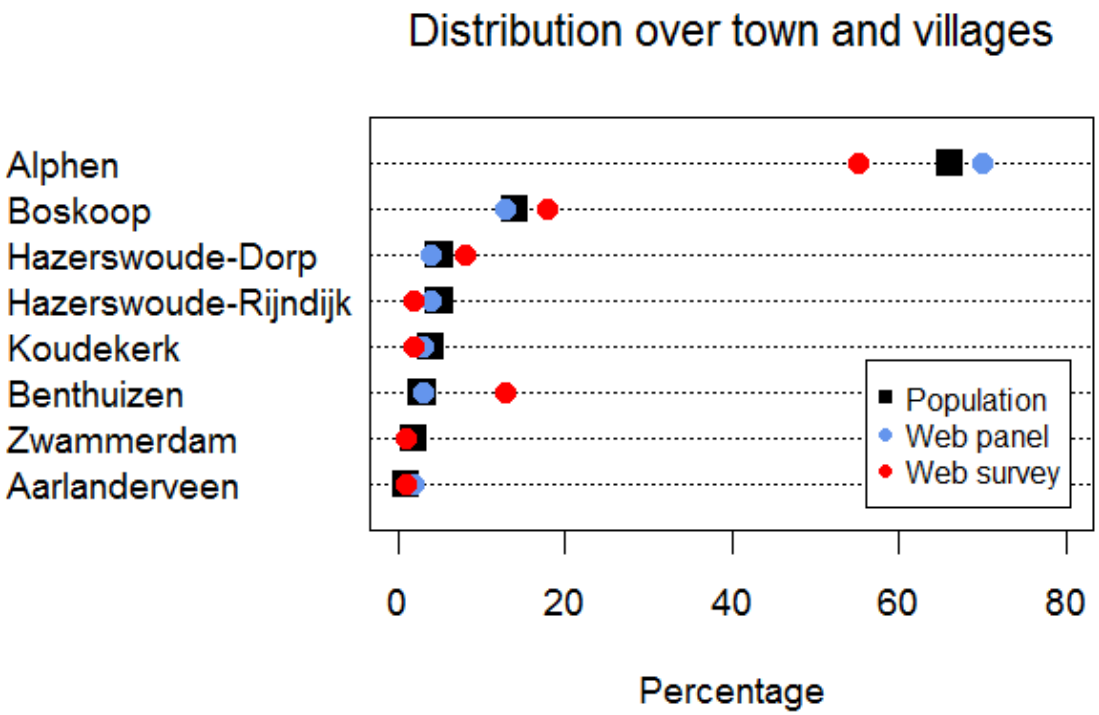
## Note

- Appeal by churches to their members to vote.

# Example 1: Shopping Sundays

## Shopping Sundays in Alphen a/d Rijn

- Distribution of the response over town and villages.

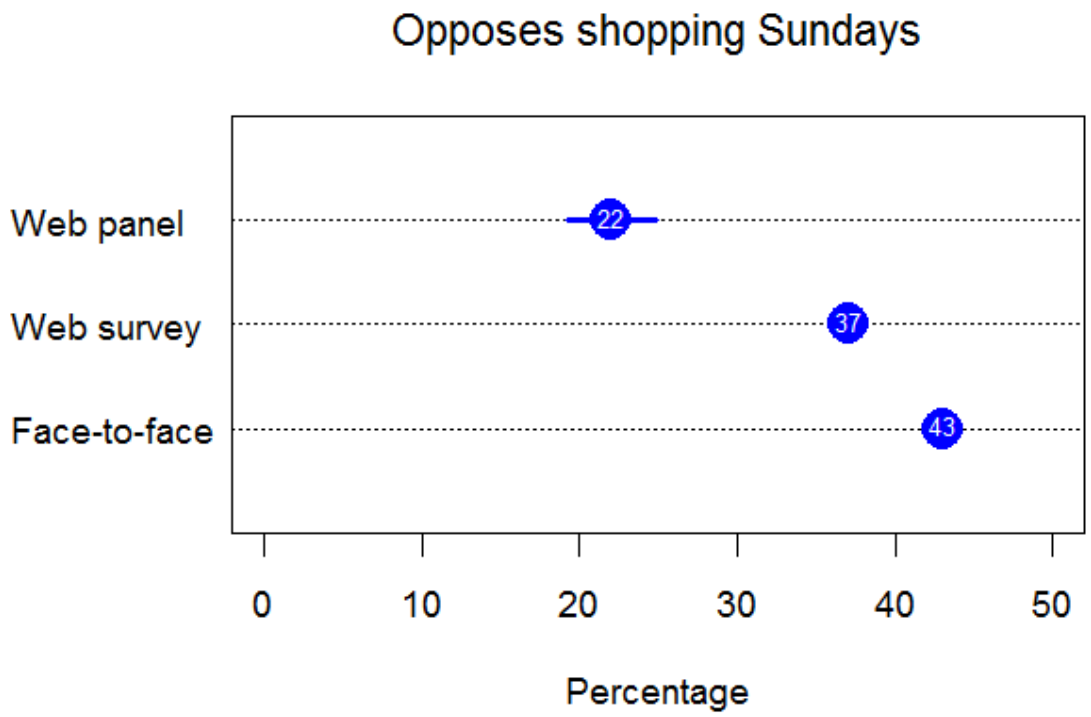



- Small Christian villages are over-represented

# Example 1: Shopping Sundays

## Shopping Sundays in Alphen a/d Rijn

- Results of the surveys

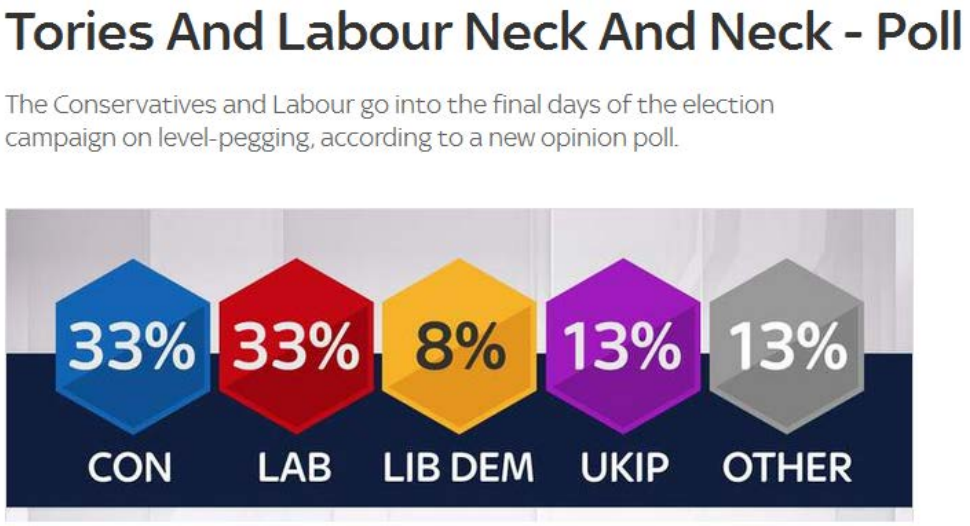



- Large differences between surveys!. Which one is correct?

# Example 2: The UK Polling Disaster

## Polls

- General election of 7 May 2015 in the United Kingdom.
- There were many polls.
- All predicted a neck-to-neck race between the Conservative Party and the Labour Party, likely leading to a 'hung parliament'.
- They were all wrong: the Conservatives got a comfortable majority of 6.5 percentage points.




*Sky News*
*4 May 2015*

# Example 2: The UK Polling Disaster

## Polls

- Difference between Conservatives and Labour.

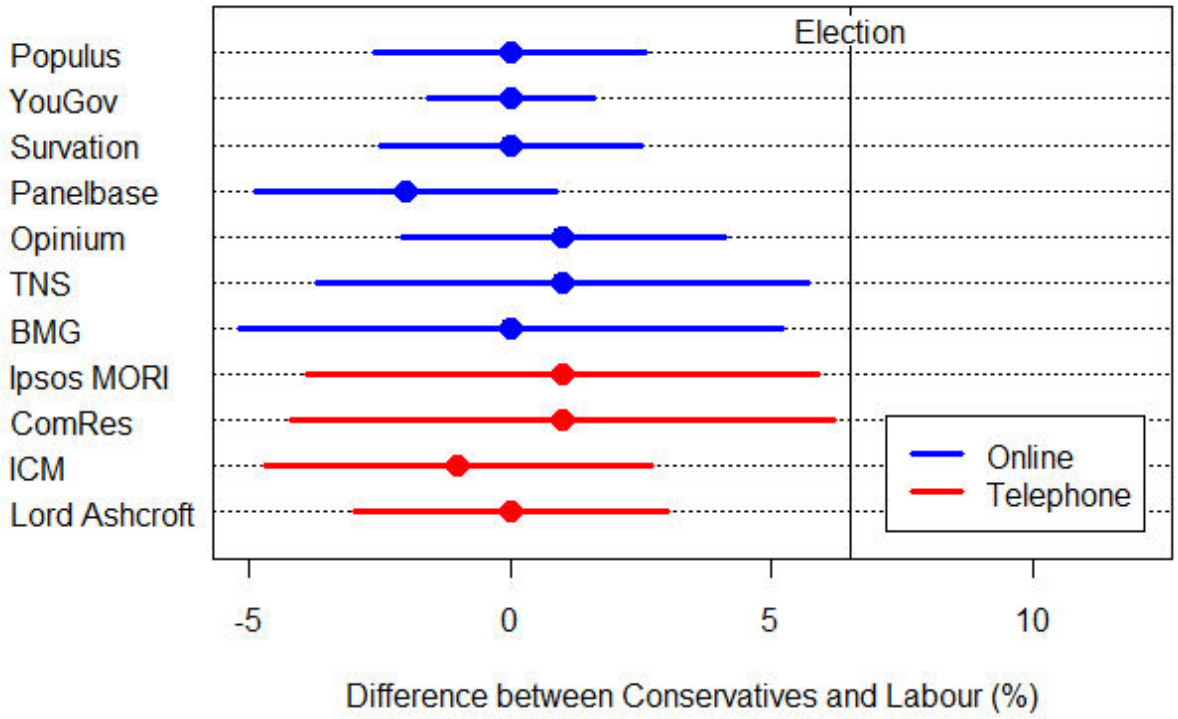
| Poll | Mode | Sample | Difference |
|---|---|---|---|
| Populus | web panel | 3,917 | 0% |
| YouGov | web panel | 10,307 | 0% |
| Survation | web panel | 4,088 | 0% |
| PanelBase | web panel | 3,019 | -2% |
| Opinium | web panel | 2,916 | 1% |
| TNS | web panel | 1,185 | 1% |
| BMG | web panel | 1,009 | 0% |
| Ipsos MORI | telephone | 1,186 | 1% |
| ComRes | telephone | 1,007 | 1% |
| ICM | telephone | 2,023 | -1% |
| Lord Ashcroft | telephone | 3,028 | 0% |

- Difference in election: 6.5%

# Example 2: The UK Polling Disaster

## Polls

- Variable: difference between Conservatives and Labour.




- There are significant differences between polls and election result.

# Example 2: The UK Polling Disaster

## Investigation by British Polling Council

- There was no 'Shy Tory Factor'.
- There was no 'Late Swing'.
- 'Herding' could not be excluded completely.

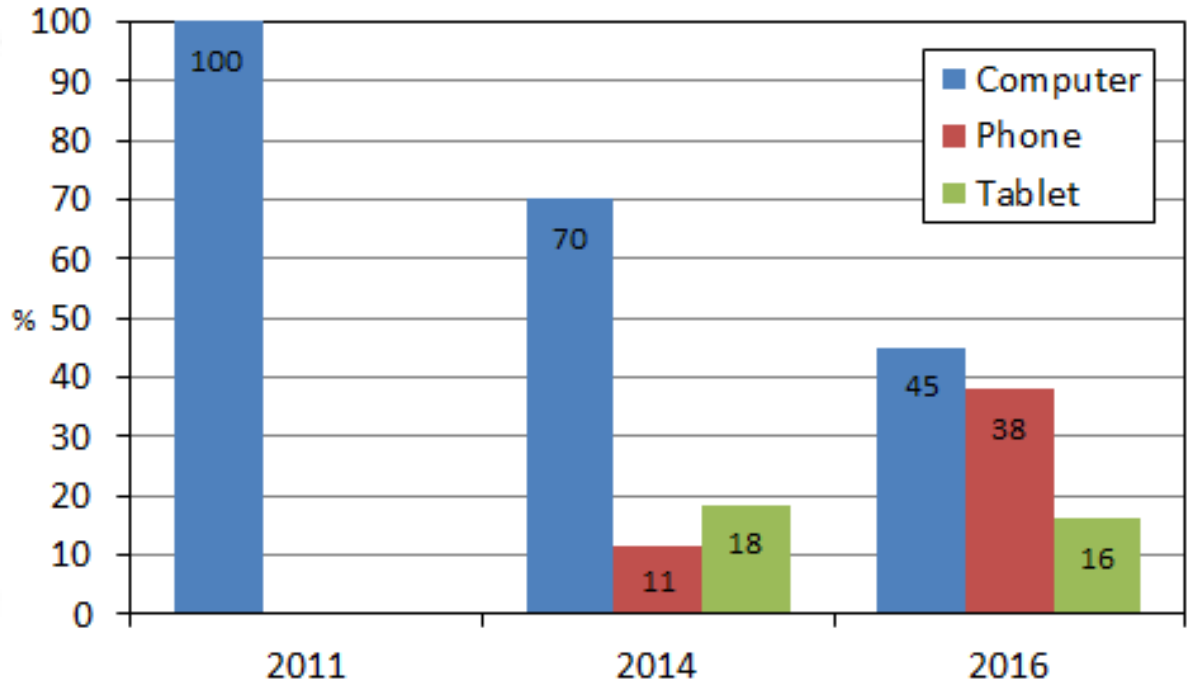
## Conclusions

- The web surveys were not representative because they were based on self-selection web panels.
- The telephone surveys were not representative because they suffered from very low response rates (20%).
- The weighting adjustment techniques used, were not effective. They were not able to reduce the bias.

# Future development

## New internet devices

- From computer to smartphone.
- Website of the Dutch Automobile Association, 10 million hits/month in 2016.



# Some conclusions

## Online data collection

- Online data collection will become more and more important.
- The outcomes of online surveys will only be accurate if the sample is selected using probability sampling.
- Be aware of self-selection surveys.
- There is no guarantee that correction (adjustment weighting) will reduce the bias of estimates.

## Issues

- Survey costs. Quality has its price.
- Recruitment by self-selection.
- Low response rates.
- A large sample does not always mean a better sample.