Challenges of web surveys and web panels

Jelke Bethlehem

Leiden University, the Netherlands

Survey data collection through the ages

Traditional data collection

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

Computer-assisted interviewing

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

And now …

- *Web surveys* (and *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.
- *Accuracy* of estimators can be computed (confidence intervals).

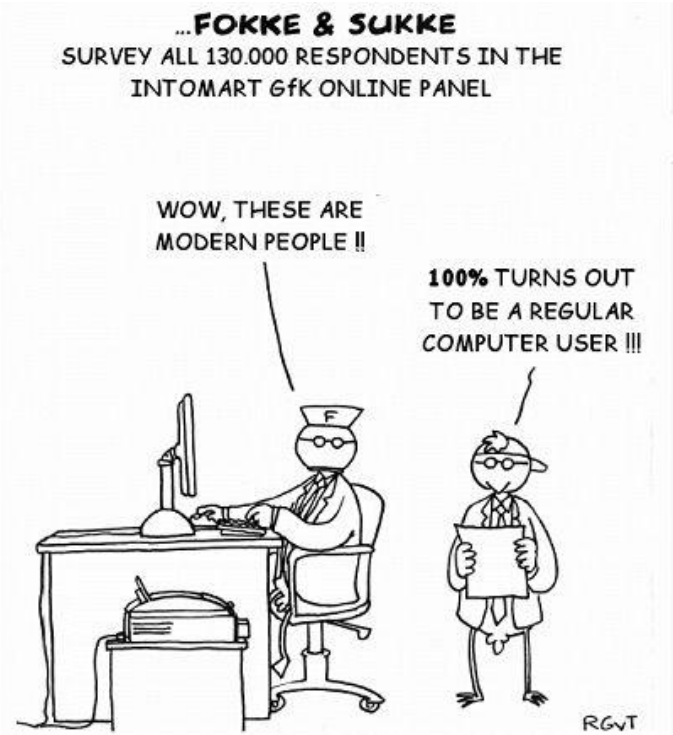
Warning

- For other forms of sampling (e.g. quota sampling), it is not clear how reliable and accurate the outcomes are.

The challenges of web surveys

Overview

- The rise of web surveys
- Under-coverage problems
- Self-selection problems
- Nonresponse problems
- Bias correction
- Measurement problems
- Web panels



The challenges of web surveys and web panels

Why are web surveys so attractive?

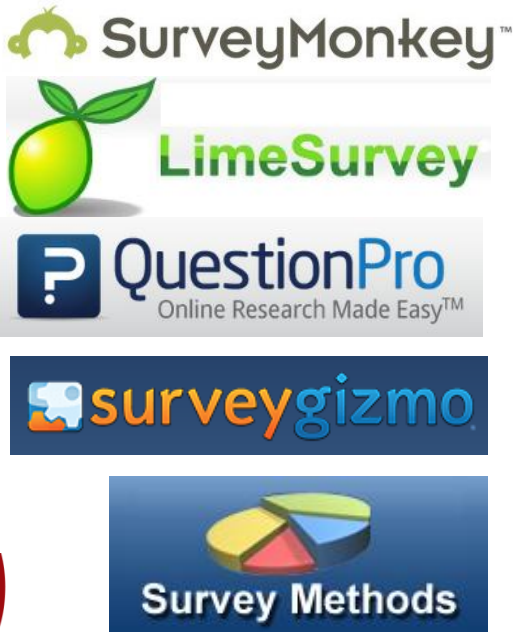
- Easy: simple access to large group of potential respondents.
- Cheap: no interviewers, no printing, no mailing.
- Fast: a poll can be launched very quickly.
- Everybody can do it!

The methodological challenges

- Under-coverage.
- Sample selection problems.
- Nonresponse.
- Measurement errors.

The question

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



Under-coverage problems

Under-coverage

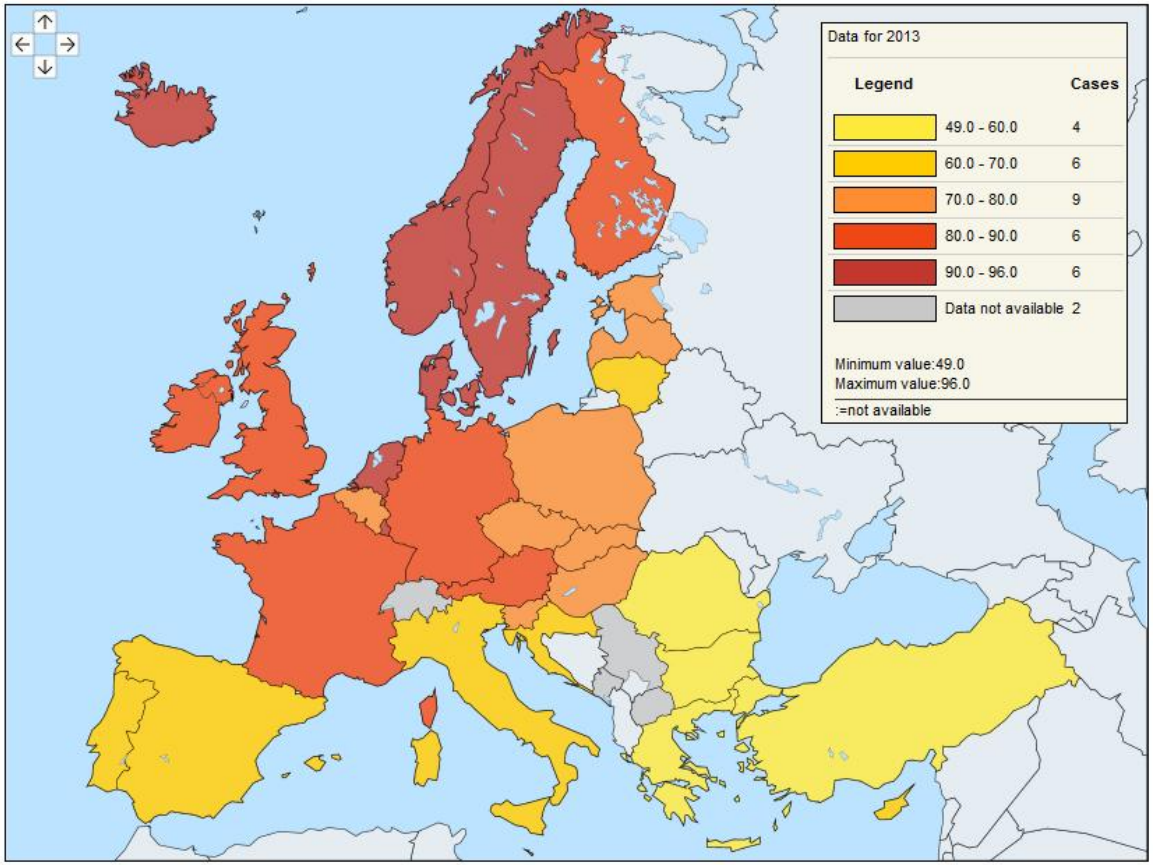
- The target population of a web survey is often much wider than those having access to the internet.
- Those without internet may differ from those with Internet.
- People without internet will never be selected for the survey.
- Therefore, estimates based on web surveys are often biased.

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 e-mail address.

Under-coverage problems

Internet-coverage in Europe in 2013



Top 3:
Iceland (96%)
Netherlands (95%)
Norway (94%)

Bottom 3:
Greece (56%)
Bulgaria (54%)
Turkey (49%)

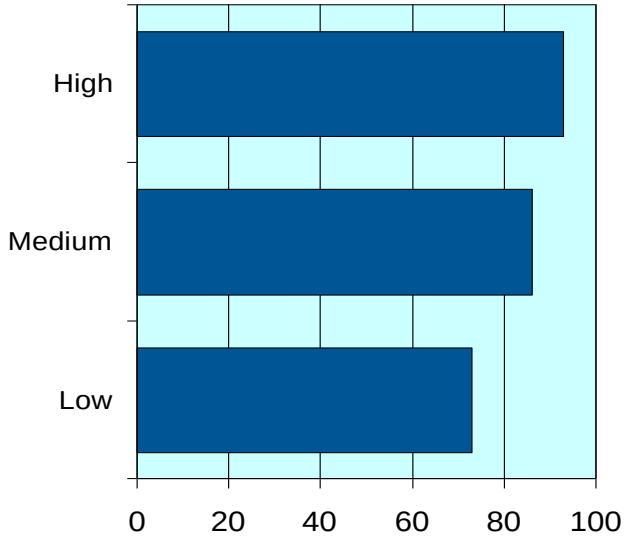
Source: Eurostat

Under-coverage problems

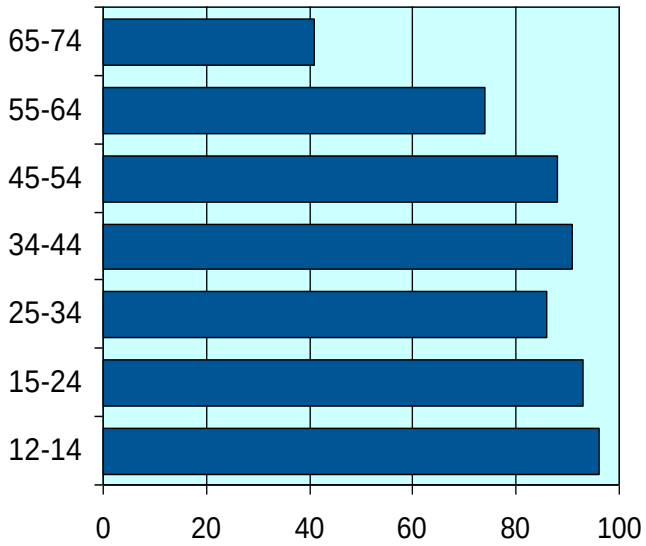
Groups with lower internet coverage

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

Coverage by education



Coverage by age



Under-coverage problems

Notation

- The target population is divided into an *internet-population* of size N_I and a *non-internet-population* of size N_{NI}.
- Internet indicator I_k:
  - $I_k = 1$ if person k has access to internet.
  - $I_k = 0$, otherwise

Means for the target variable Y

- Target population: $$\bar{Y} = \frac{Y_1 + Y_2 + ... + Y_N}{N}$$
- Internet-population: $$\bar{Y}_I = \frac{I_1 Y_1 + I_2 Y_2 + ... + I_N Y_N}{N_I}$$
- Non-internet-population: $$\bar{Y}_{NI} = \frac{(1 - I_1) Y_1 + (1 - I_2) Y_2 + ... + (1 - I_N) Y_N}{N_{NI}}$$

Under-coverage problems

Simple random sample (without replacement)

- Set of indicators a_1, a_2, .., a_N.
  a_k = 1 if element k is selected. a_k = 0 otherwise.
- Sample mean: $$\bar{y}_I = \frac{a_1 I_1 Y_1 + a_2 I_2 Y_2 + \ldots + a_N I_N Y_N}{n_I}$$
- Bias: $$B(\bar{y}_I) = E(\bar{y}_I) - \bar{Y} = \bar{Y}_I - \bar{Y} = \frac{N_{NI}}{N}(\bar{Y}_I - \bar{Y}_{NI})$$

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.

Conclusion

- 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 (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 you cannot send an e-mail with a link to questionnaire website.

Alternatives

- Draw a random sample from a population register, and send a letter (with a link) to each selected person.
- 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?

- No probability sampling is applied.
- Participants are people that have internet, happen to see the invitation, and decide to participate.
- It is a cheap and fast way to collect a lot of data.
- However, the sample is not representative.

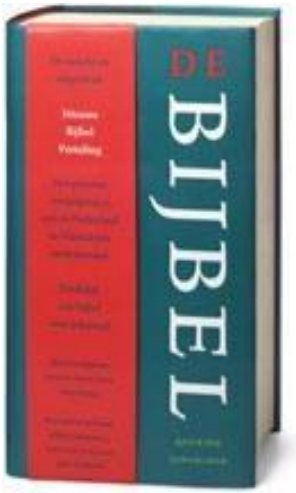
Problems

- 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 1

- 2005 Book of the Year Award.
- High profile literary prize in the Netherlands.
- A web survey was carried out to select the best book.
- One could vote for one of the nominated books or suggest another book.
- Number of participants: 90,000.
- 72% voted for a non-nominated book: the new interconfessional Bible translation.
- Result of a campaign by Bible societies, a Christian TV-channel, and a Christian newspaper.



Self-selection problems

Example 2

- 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

Example 3

- Major political poll, based on self-selection web panel.
- Attempt in 2012 to create 2,500 fake panel members.
- They would first vote for Christian-democrats (CDA) and then change to the elderly party 50PLUS.
- Attempt was discovered by the poll organisation: too many people tried to join the panel at the same day.

Self-selection problems

Example 4, election polls, 2012, seats in parliament

- Major opinion polls, one day before the parliamentary election.
- All polls based on self-selection web panels

Party	Election result	Peil.nl	Politieke Barometer	TNS NIPO	De Stemming
VVD	41	36	37	35	35
PvdA	38	36	36	34	34
PVV	15	18	17	17	17
CDA	13	12	13	12	12
SP	15	20	21	21	22
D66	12	11	10	13	11
GL	4	4	4	4	4
CU	5	5	5	6	7
SGP	3	3	2	2	3
PvdD	2	3	3	2	2
50PLUS	2	2	2	4	3
Tot. diff.		18	18	24	24
Mean diff.		1.6	1.6	2.2	2.2

Self-selection problems

Self-selection sample

- No probability sampling. Participants are people that have internet, happen to see the invitation, and decide to participate.
- Participation indicators R_1, R_2, …, R_N, each with value 1 or 0.
- Each person k has unknown probability π_k to participate, with $P(R_k = 1) = \pi_k$.
- Sample size $n_S = R_1 + R_2 + \ldots + R_N$.
- Sample mean: $\bar{y}_S = \frac{1}{n_S} \sum_{k=1}^{N} R_k Y_k$
- Expected value: $E(\bar{y}_S) = \frac{1}{N\bar{\pi}} \sum_{k=1}^{N} \pi_k Y_k$
- Bias: $B(\bar{y}_S) = \frac{R_{\pi Y} S_\pi S_Y}{\bar{\pi}}$ ($R_{\rho Y}$ =correlation)

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. The smaller the participation probabilities, the larger the bias.
- The strength of the relationship between participation behaviour and the target variable of the survey. The stronger the correlation, the larger the bias
- The variation in the participation probabilities. The larger the variation, the larger the bias.
- Note: the expression is similar to that of the bias due to nonresponse. However, the participation probabilities are much smaller. So the bias is much larger.

Self-selection 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!

Nonresponse problems

The nonresponse problem

- 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, often not more than 30%.

Nonresponse problems

Probability sample

- Simple random sample: set of indicators a_1, a_2, .., a_N.
  a_k = 1 if selected, and a_k = 0 if not selected.
- Each person k has unknown probability ρ_k to respond.
- Response indicators R_1, R_2, …, R_N, R_k=1 (response) or 0 (nonresponse)
- So $P(R_k = 1) = \rho_k$.
- Sample size $n_R = a_1R_1 + a_2R_2 + \ldots + a_NR_N$.
- Sample mean: $\bar{y}_R = \frac{1}{n_R}\sum_{k=1}^{N} a_k R_k Y_k$
- Expected value: $E(\bar{y}_R) = \frac{1}{N\bar{\rho}}\sum_{k=1}^{N} \rho_k Y_k$
- Bias: $B(\bar{y}_R) = \frac{R_{\rho Y} S_\rho S_Y}{\bar{\rho}}$ ($R_{\rho Y}$ =correlation)

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. The smaller the response probabilities ρ_1, ρ_2, …, ρ_N, the larger the bias.
- The strength of the relationship between response behaviour and the target variable of the survey. The stronger the correlation, the larger the bias.
- The variation in the response probabilities. The larger the variation, the larger the bias.

Representativity problems

Bias correction

- Apply *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.

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.

Measurement problems

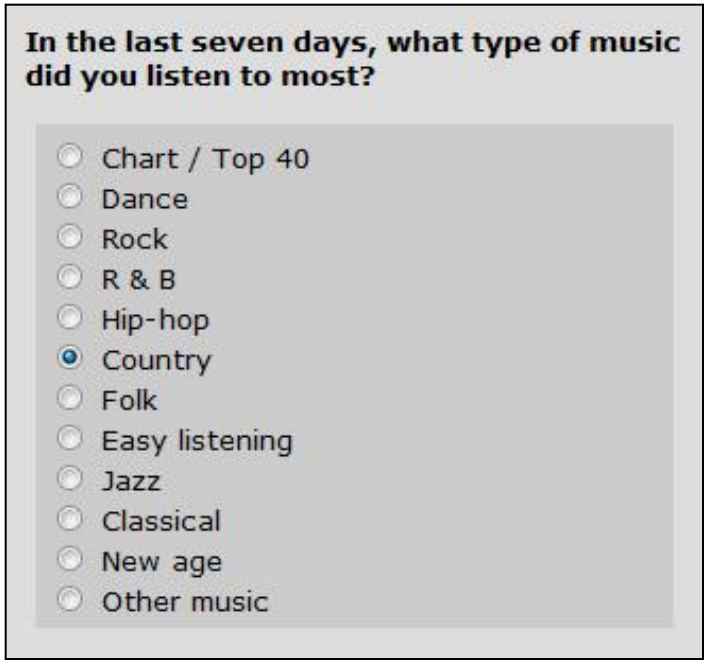
Online data collection

- There are no interviewers. Respondents are on their own.
- Respondents are not really interested. Participation is not important.
- They do not read the question, but just scan the text.
- *Satisficing*: they do not select the optimal answer, but the first reasonable answer.
- There is no penalty for wrong answers.
- Bother respondents with error messages?
- Enforce routing?

Measurement problems

Satisficing

- *Primacy effect*: preference for answer early in the list.

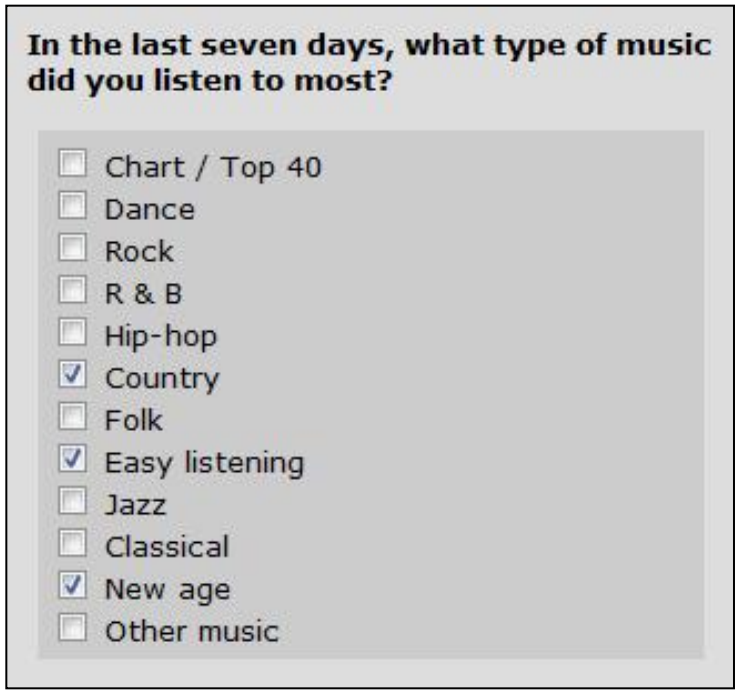


- *Endorsing the status quo*: selecting the no-change answer.
- *Preference for middle (neutral) answer* in opinion question.

Measurement problems

Satisficing

- Preference for *don't know*.
- *Arbitrary answers* for check-all-that-apply questions



Measurement problems

Satisficing

- *Straight-lining* in matrix questions.

	Excellent	Very good	Good	Fair	Poor
1. How would you rate the overall quality of the radio station?	○	○	◉	○	○
2. How would you rate the quality of the news programs?	○	○	◉	○	○
3. How would you rate the quality of the sport programs?	○	○	◉	○	○
4. How would you rate the quality of the music programs?	○	○	◉	○	○

Possible problems in CAPI/CATI surveys

- *Acquiescence*: tendency to agree with statements in questions.
- More socially desirable answers to *sensitive questions*.

Measurement problems

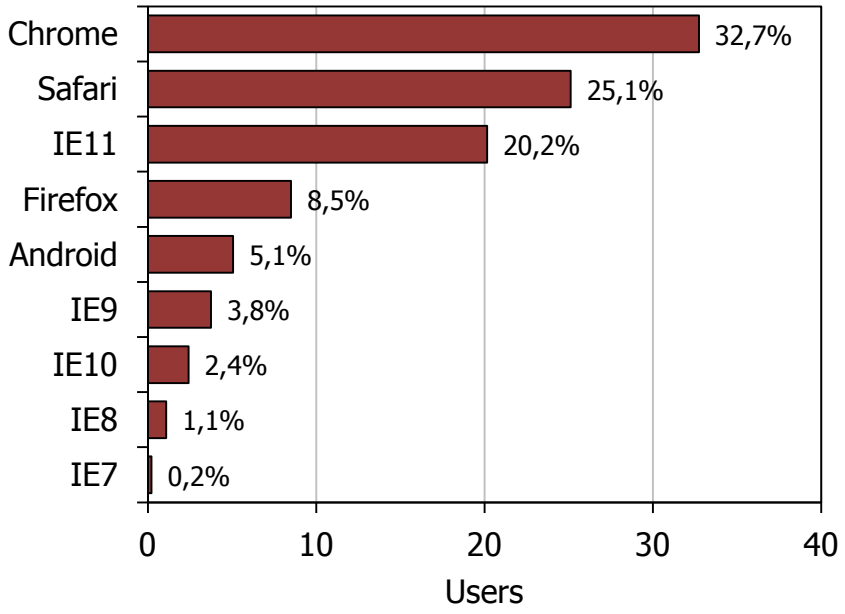
Technical challenges

- Questionnaire must look the same in all versions of the same browser.
- Questionnaire must look the same in all browsers that are used (Internet Explorer, Firefox, Chrome, Safari, etc).
- Special features (Javascript, Flash, SVG) may not work.
- Screen size unclear.
- Browser may be opened in a small window.
- Different devices: PC, laptop, netbook, tablet, smartphone.
- More and more, people check their e-mail on a smartphone. So the survey must run on a smartphone.

Measurement problems

Technical challenges

- More than 50% of the internet users read their mail on a tablet or a smartphone. So web surveys must work on these devices.
- Users of the ANWB website (Automobile Association), 8 million visitors, May 2015:



Safari: mostly tablet/phone.
Chrome: partly tablet/phone.

Web panels

What is a web panel

- People are recruited only once.
- They participate several times in surveys taken from the panel.

Why a web panel?

- Instrument for longitudinal research.
- Sampling frame for cross-sectional research.
- Quick ad-hoc surveys.

Additional challenges

- How to recruit a representative online panel?
- Nonresponse (in recruitment and surveys).
- Maintenance (attrition, panel conditioning).

Web panel recruitment (1)

Recruitment by means of self-selection (opt-in)

- People decide themselves whether or not to become a member of the panel. No probability sample selection.
- Participation probabilities are unknown. This leads to biased estimates.

Other self-selection problems

- Also people from outside the target population can become a member of the panel.
- Sometimes multiple membership is possible.
- Groups of people may attempt to manipulate the outcomes of the polls.

Conclusion

- A self-selection panel is out of the question for general population surveys.

Web panel recruitment (2)

Recruitment by means of probability sampling

- Allows for unbiased estimation.
- Allows for computation of margins of error.
- Required: a sampling frame with e-mail addresses.
- Such a sampling frame is not available.
- Solution: different mode(s) for recruitment: mail, CATI or CAPI (or a combination).
- Traditional sampling frames can be used.
- Disadvantage: recruitment is time-consuming and expensive.

Conclusion

- Probability sampling is ok, but it is time-consuming and expensive.

Web panel recruitment (3)

Recruitment from other surveys

- Build panel from respondents of previous CAPI or CATI surveys.
- Respondents may have agreed to participate in future surveys.
- Recruitment may be less expensive.
- But these respondents may be a selective group, and therefore the resulting panel may lack representativity.

Conclusion

- Probably not a good idea.

Nonresponse in web panels

Nonresponse 1: recruitment nonresponse

- High, as participation requires substantial commitment.
- Bias reduction (adjustment weighting) difficult due to lack of relevant auxiliary variables.

Nonresponse 2: survey/wave nonresponse (attrition)

- May be low, as people agreed to participate.
- Plenty of auxiliary variables for bias reduction, e.g. from *profile survey*.

Some final remarks

Can you use web surveys and web panels in a scientifically sound way?

- Yes, if the sample is selected by means of probability sampling.
- No, if the sample is selected by means of self-selection.
- Yes, if the under-coverage problem is solved, for example by using mixed-mode surveys, or giving people internet access.
- No, if you want high response rates.
- Yes, if you just want to reduce costs and you solve the problems caused by mode effects.

- Can you cope with nonresponse bias?
- Can you cope with measurement errors?
- Can you cope with technical issues?