methods, data, analyses | Vol. 8(1), 2014, pp. 5-24 DOI: 10.12758/mda.2014.001

The Accuracy of Pre-Election Polling of German General Elections

Rainer Schnell, Marcel Noack
University of Duisburg-Essen

Abstract

Pre-election polls are the most prominent type of surveys. As with any other survey, estimates are only of interest if they do not deviate significantly from the true state of nature. Even though pre-election polls in Germany as well as in other countries repeatedly show noticeably inaccurate results, their failure appears to be quickly forgotten.

No comparison considering all available German data on actual election results and the confidence intervals based on pre-election polls has been published. In the study reported here only 69% of confidence intervals covered the election result, whereas statistically 95% would have to be expected. German pre-election polls even just a month ahead are therefore much less accurate than most introductory statistical textbooks would suggest.

Keywords: Pre-Election-Polls, Empirical coverage, Confidence intervals for binomial data, Design effects, Sonntagsfrage

1 Introduction

Pre-election polls account for a large proportion of political media coverage. The interest in election forecasts is based on the assumption that election results can be precisely predicted (Crespi 1988: 4). Contrary to this assumption, the available literature records a long series of failures that is not limited to either specific countries, or time periods. Common examples are the American presidential elections in 1948 and 1996 (Mitofsky 1998), the election of the British House of Commons 1992 (Lynn/Jowell 1996: 22), the French presidential election in 2002 (Durand et al. 2004), the Italian parliamentary election 2006 (Callegaro/Gasperoni 2008) and the 2005 Bundestag election (Groß 2010: 9).[1] However, the fact of its repeated failure does not appear to be common knowledge. For example, several contemporary German textbooks of statistics present naïve and uncommented calculations of confidence intervals based on pre-election polls.[2] Those computations rely on the same erroneous assumptions on confidence intervals and their interpretation as the sometimes reported *margins of error* in media coverage of pre-election polling. All these computations ignore the additional problems of surveys on human populations in general (Groves 1989: IV) and the specific problems of pre-election polls (Wüst 2010). Since these problems introduce more uncertainty in estimates for population parameters, the accuracy of pre-election polls in Germany is much lower than the naïve margins of error computations suggest as we will show.[3]

Direct correspondence to
Rainer Schnell / Marcel Noack, University of Duisburg-Essen, Research Methodology Group, Lotharstr. 65, 47057 Duisburg, Germany
E-mail: rainer.schnell@uni-due.de / marcel.noack@uni-due.de

1 Research on the development of election forecasts over time is available for some countries. For Portugal 1991-2004, see Magalhães (2005); for the United Kingdom 1950-1997, see Sanders (2003); for the USA 1979-1987, see Crespi (1988); for Germany 1947-2009, see Groß (2010).

2 For example Behnke et al. (2006: 397-399), Bosch (2012: 180-181), Fahrmeir et al. (2007: 393), Gehring/Weins (2009: 266-268), Klammer (2005: 124), Luderer (2008: 98) or Oestreich/Romberg (2012: 243-244).

3 This discrepancy between textbooks and empirical facts is hard to explain. One possible mechanism is due to the ambiguity of the German word *Wahlprognose*. The international scientific literature distinguishes between exit polls and pre-election polls. In German, the words *Wahlprognose* and *Hochrechnung* are used for both kinds of surveys. Since the high level of precision of exit polls in Germany leaves no room for further improvement (Hilmer 2009: 258), this accuracy is probably falsely attributed to all kinds of election polls.

In the following, we present a comprehensive statistical review on the performance of German pre-election polls of general elections between 1957 and 2013 based on specific voting intentions (*Sonntagsfragen*).[4]

2 Methodological Problems of Pre-election Polls

The purpose of any survey is the estimate of a population parameter μ by a sample statistic $\hat{\mu}$. In this context, the central concept is the *Total Survey Error* model (TSE). The most commonly used criterion of quality within the TSE is the *Mean Squared Error* (MSE),

$$\text{MSE}(\hat{\mu}) = \text{Bias}^2 + \text{Variance} \tag{1}$$

which is the sum of the squared Bias (difference between the expectation of the estimate $\text{E}(\hat{\mu})$ and the population parameter μ) and the variance of the estimate (Schnell 2012: 387).

The main sources of error for the MSE are specification error, frame error and non-response error on the side of bias; sampling mostly affects the variance of the estimate. Measurement errors and data processing errors are equally relevant for both bias and variance (Biemer/Lyberg 2003: 59).

Any of these error sources can have such a severe impact that any conclusions drawn from the data have to be considered as false (Alwin 2007: 3). Therefore the objective of a good survey design is to minimize the sources of these errors, taking into account the available resources and other limiting factors (Biemer/Lyberg 2003: 38). For this reason, detailed information on the design and execution of a survey are essential in order to assess its quality.

The mode of sampling is of central importance for the errors of surveys. Hence, the methodological literature on pre-election polls agrees that quota sampling should be avoided (Lynn/Jowell 1996). With the exception of the IfD Allensbach, quota samples are therefore rarely used in Germany. In accordance with recommendations of the *Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute* (ADM) in 1979, random samples today constitute the norm for such election forecasts (Groß 2010: 49). Those are mostly CATI interviews via random digit dialing. Since the 1950s, the response rate in academic studies has nearly halved (Schnell 2012: 164). The low response rate is due to a decrease in both cooperation

4 In the German political science literature, a distinction beween opinion, projection and prognosis (*Stimmung*, *Projektion*, *Prognose)* has been introduced by Wüst (2003). The first two constructs reflect specific voting intentions and a more or less theoretical weighting of the opinion. Prognosis is reserved for exit polls. However, most publications of polling institutes and German political scientists refer to pre-election poll results based on specific voting intentions as prognosis.

and availability of respondents. Reducing the number of non-contacts to less than 5% requires long field periods as well as a high number of attempts of contact. Pre-election polls do not necessarily implement either. The German television *Politbarometer*, for instance, operates on a field period of four days (Schnell 1997: 117).

Few of the publications of pre-election poll results include information considered as necessary by the professional *standards for disclosure* (AAPOR 2010, similar: ICC/ESOMAR 2008): information on sponsor and surveying institute, the exact phrasing of questions and response categories, details on the sampling frame and problems of coverage, mode of sampling, sample size, standard error, type of weighting, design effects, instructions to the interviewers, notification letters, screening procedures, incentives, detailed information on response rates as defined by AAPOR, interviewer training and interviewer workload and assignment.[5]

Another problem in pre-election polling is known as *political weighting* of the raw data. No algorithm for the computation of these correction factors has been published (Groß 2010: 110). Using Politbarometer-data, Groß (2010: 110) estimated the impact of political weighting on published results between 1986 and 2005. He showed a small mean difference between published and raw data, but a considerable variance of this difference. Further methodological details required for the evaluation of survey results are withheld by the institutes. Information on response rates, contact strategies of the interviewers, or sampling design are reported rarely, or not at all (Groß 2010: 109-111).[6]

Further technical details will nearly always be missing. This includes, for example, the strategy of dealing with hard-to-reach respondents, whose voting behavior can differ from that of easy-to-reach respondents (Crespi 1988: 43). Ever since the "Literary Digest Disaster" of the US election in 1936, problems of coverage and selective non-response bias have been discussed in the methodological literature as the possible causes of failure of election forecasts (Lusinchi 2012; Walsh et al. 2009: 317; Frankovic et al. 2009: 575-587).

Furthermore, individuals who are only available via mobile phone might cause sampling problems.[7] Even when they had a positive and known selection probabil-

---

5 Paragraph 11b of the ESOMAR standards states: "Where any of the findings of a research project are published by the client, the latter shall be asked to consult with the researcher as to the form and content of publication of the findings. Both the client and the researcher have a responsibility to ensure that published results are not misleading" (ICC/ESOMAR 2008). Infratest and Emnid are institutional members of ESOMAR, whereas only some people working for Forsa and Allensbach are members. As a consequence Forsa and Allensbach are not bound by the guidelines. By comparison, the ADM-standards are less mandatory (ADM 1999). As opposed to ESOMAR standards, ADM institutes are not factually responsible for publications of the sponsor. Although the required details as mentioned in the AAPOR standards could easily be published on the pages of the ADM or the institutes, this rarely happens.

6 Walsh et al. (2009: 317) report the same for the USA.

7 On the consequences of these so-called "cell phone onlys", see AAPOR (2009: 31).

ity, the interview situations still cannot be compared and reported voting behavior might differ between respondents on landlines and on mobile phones. Information on response rates distinguishing between mobile phone and landline numbers in German pre-election polling is rarely published.

In general, systematic differences between respondents and non-respondents will cause biased estimates.[8] Therefore, the exclusion of very small subgroups can have a high impact on the results. The details needed to estimate these effects are unfortunately hardly ever reported in the case of election coverage by means of pre-election polls. Since the technical details needed for a methodological analysis of a pre-election poll are seldom published, currently no comprehensive methodological analysis of pre-election polls is possible in Germany.[9] This paper will therefore be limited to a statistical analysis of the quality of pre-election polling as forecasting method.

3 Data

The following analyses are based on a dataset of a total of 232 published pre-election polls on the German general elections between 1957 and 2013. This dataset is

8 This non-response bias is given via

$$\bar{y}_{\text{Res}} - \bar{y}_{\text{All}} = \frac{n_{\text{Non}}}{n_{\text{Non}} + n_{\text{Res}}}(\bar{y}_{\text{Res}} - \bar{y}_{\text{Non}})$$

with the respective values for all respondents, respondents (Res) and non-respondents (Non) (for an example, see Groves (1989: 134)). One possibility of estimating the maximum bias would be via the response propensities ρ using the R indicator approach $R(\rho)$ with

$$B_{\max}(y,\rho) = \frac{(1 - R(\rho))S(y)}{2\bar{\rho}} \geq \left|\frac{\text{Cov}(y,\rho)}{\bar{\rho}}\right|$$

with

$$R(\rho) = 1 - 2S(\rho)$$

given, where $S(\rho)$ is the standard deviation of the response propensities, $S(y)$ the variance of the dependent variable in the population, $\bar{\rho}$ the mean of the response propensities and $\text{Cov}(y, \rho)$ the covariance of the response propensities and the dependent variable. The probability ρ that a sampled individual actually answers, is estimated with a number of auxiliary variables x_j, for example, via logistic regression (Schouten et al. 2009: 105). The bias is greater, the stronger the correlation between response propensities ρ and the variable of interest y (Schouten et al., 2009: 107). However, it has to be noted that the selection of the auxiliary variables x_j is of great importance. If the non-response mechanism does not correlate with the auxiliary variables used to estimate the response propensities, the bias will remain unnoticed (Schnell, 2012: 174). Using irrelevant auxiliary variables will miss any existing bias.

9 In the US, the work of Crespi (1988) is still the most extensive methodological analysis of pre-election polls. Recent minor additions can be found in Lau (1994) and DeSart/Holbrook (2003), as well as in Keeter et al. (2000) and Keeter et al. (2006).

a subset of a dataset provided by Groß containing 3610 polling results published between 1949 and 2009 (Groß 2010: 121-126). To reduce the chance of potential last minute swings, only pre-election polls with a sufficiently small temporal distance between poll and election were used. Because of that, polls published more than one month before an election were excluded.[10]

Sample size is a necessary information for the computation of standard errors and confidence intervals. For the majority of the remaining 204 studies, this critical information was not included (n=108). Through extensive archival research, sample size for additional 84 of those prognoses could be determined.

Most of the remaining 24 studies were older than 25 years. Hence, no further details on the studies could be found.[11] For these studies, we used the median sample size of the studies before 1990 (n=1000). Since they met the inclusion criteria of our study, we appended 28 recent polls covering the general election in 2013 to the dataset.[12]

At least 19% of the polls are based on quota samples.[13] Quota samples are no probability samples, therefore inference for quota samples cannot be justified statistically. Pre-election polls based on quota samples are only treated as random samples for the purpose of comparison.

4 Methods

Survey estimates should be reported together with their corresponding confidence intervals (CI). The precision of the estimation is given by the width of the CI.[14] The narrower the CI, the more precise the estimate. If every possible sample, of fixed size, is drawn from the same sampling frame, and a CI is calculated for each independent sample, a well-defined proportion of CIs contains the true parameter. That well-defined proportion is called the coverage probability or confidence level (Särndal et al. 1992: 55). If all statistical assumptions required for the calculation of

10 On request, the original dataset was kindly provided by Jochen Groß.

11 For 15 of these studies, the publications also do not mention the polling companies, which greatly complicates the research.

12 The data were extracted from the web page: www.wahlrecht.de.

13 This is not always apparent from the publications. Since 19% of the polls have been published by a German company which nearly always uses quote samples (namely Allensbach) 19% quota samples is a conservative estimate.

14 Another approach would be the usage of prediction intervals (for a review see Krishnamoorthy/Peng 2011). The difference between those two types of intervals is the intended use. Prediction intervals try to predict a future observation (Devore/Berk 2012: 404). Confidence intervals are statements on the uncertainty of population parameter estimates. Therefore, prediction intervals are not appropriate for our kind of analysis. Of interest here is the latter kind of inference.

CIs are met, a CI can accurately be determined analytically, that is without drawing all possible samples.

Assuming the election results is the population parameter, it can be checked whether the parameter is contained within the corresponding CIs. The number of CIs containing the parameter can be counted. If the assumptions are met, the proportion of CIs containing the parameter should be equal to the coverage probability. If reports of polling results mention sampling errors at all, they almost always report CIs for simple random samples, assuming a binomial distribution.[15] Statistically, this is erroneous in several respects.

Pre-election polls in Germany are hardly ever based on simple random samples, but on complex sampling designs. Nearly always, a complex design will result in a higher standard error than a simple random sample of the same size (Schnell 1997: 272-284). There are essentially two causes for the loss of precision. First of all, most complex samples are cluster samples, so that the population is divided into disjunctive units (areas, schools, number blocks in CATI) before sampling. From each unit, a number of persons, or all, are drawn. However, individuals in a spatial unit tend to be more similar to each other, than individuals chosen independently from the population. This homogeneity within the cluster needs to be taken into account for the estimation.[16]

Furthermore, interviewers generally conduct several interviews. Given that interviews conducted by one particular interviewer are more similar than interviews conducted by different interviewers, these homogeneities cause additional loss of precision (Schnell 1997; O'Muircheartaigh/Campanelli, 1998; Schnell/Kreuter, 2005). This effect increases with the number of interviews per interviewer. Since the number of interviews per interviewer is especially high for CATI surveys, this effect is particularly strong.[17] The impact of the interviewer on the variance of the estimate can be even more severe than the effect resulting from spatial clustering (Schnell/Kreuter, 2005: 401). Unfortunately, this is largely ignored when analyzing CATI surveys.

---

15 The best-known example of pre-election polling in Germany is the public-service television Politbarometer. On their homepage: http://politbarometer.zdf.de, 15.11.2013, the CI for a sample of 1250 respondents and a share of 40% of the votes is indicated as +/- 2.7%.

16 This problem was systematically discussed at first by Kish (1965: 164); an early application to pre-election polling can be found by Converse/Traugott (1986: 1095).

17 This effect (deft) is usually simply estimated with

$$deft = \sqrt{1 + \rho(b-1)}$$

(Kish, 1965: 162), where ρ is the homogeneity within the cluster (more precisely: the intraclass correlation coefficient) and b is the mean of the number of observations within the cluster.

Statistically, the loss of precision of complex designs is called the design effect (deft). Deft is defined as the ratio of the standard error of a complex sample and the standard error of a simple random sample of the same size:

$$deft = \frac{\hat{\sigma}_{\theta,\text{complex}}}{\hat{\sigma}_{\theta,\text{SRS}}} \tag{2}$$

Using estimates of deft, adjusted CIs can be calculated, which give a correct coverage probability.

The corrected intervals are wider than the usually calculated naïve 95%-CIs, by the factor deft:

$$\left[p_i - 1.96 * deft * \sqrt{\frac{p_i(1-p_i)}{n}}; p_i + 1.96 * deft * \sqrt{\frac{p_i(1-p_i)}{n}} \right] \tag{3}$$

The naïve CIs, calculated on the assumption of a simple random sample, therefore lead to believe in a higher precision than actually given.

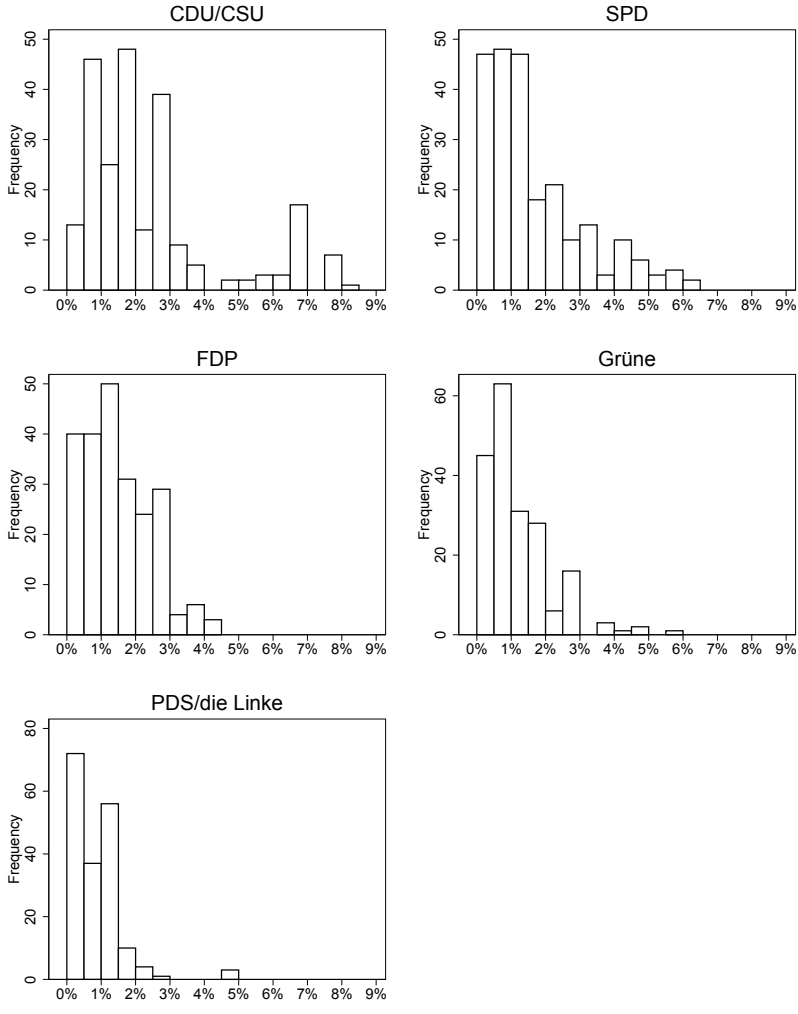
For the calculation of design effects, microdata of the variables of interest, as well as the variables that define the clusters are needed. These data are hardly ever available for pre-election polls. For this reason, an average design effect is occasionally used (UN 2005: 129). Design effects vary considerably; therefore we use the average of 118 estimations from the German Defect Project (Schnell/Kreuter, 2005: 400) with 1.4 (standard deviation=0.3) as a conservative estimate.[18] These intervals are used in the following figures.

5 Results

Of primary interest is the absolute error of the result of the pre-election result compared to the result of the general election. For each party, this is calculated as the absolute value of the difference between the survey result and the election result. Figure 1 shows the distributions of these differences. Obviously, distributions for all parties are right-skewed. Furthermore, there is a second local maximum for the CDU/CSU at 7%. This is due to the general election of 2005, when every poll mispredicted the result of the majority party (CDU/CSU). Naturally, the absolute

18 For comparison: in the Allbus 2008, questions for voting preferences for specific parties show design effects between 1.43 and 1.65 (CDU/CSU, SPD, FDP and Grüne) given the sampling point as cluster, and 1.71 to 2.03 for the interviewer as cluster. Since, as opposed to the Defect study, the Allbus 2008 is not based on an interpenetrating sampling design (Bailar 1983), the confounded effects of interviewer and sampling point cannot be separated.

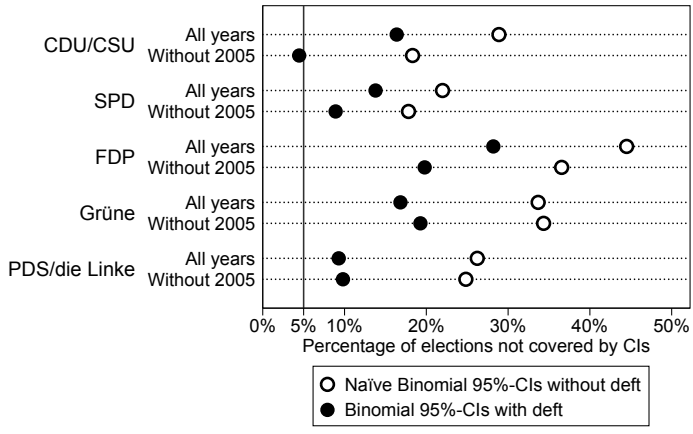
Figure 1: Absolute error of pre-election polls, 1957-2013



error for the small parties (FDP, Grüne and die Linke) is smaller than for the major parties. If the difference between prognosis and result is normed to the size of the party, the resulting relative error is considerably greater. A departure of 2% in the prediction of a party that achieved 6% corresponds to a third of its voters. 9 out of the 145 prognoses (6.2%) for parties with election results under 6% produce relative errors of this magnitude.

Please note: it is expected that at most 5% of the election results are not contained in the CIs; therefore, it is surprising that 6.2% of the poll results exceed a

Figure 2: Empirical non-coverage



third of the respective party size. The absolute error of the pre-election polls is therefore considerably greater than would be expected by a statistically naïve estimation.

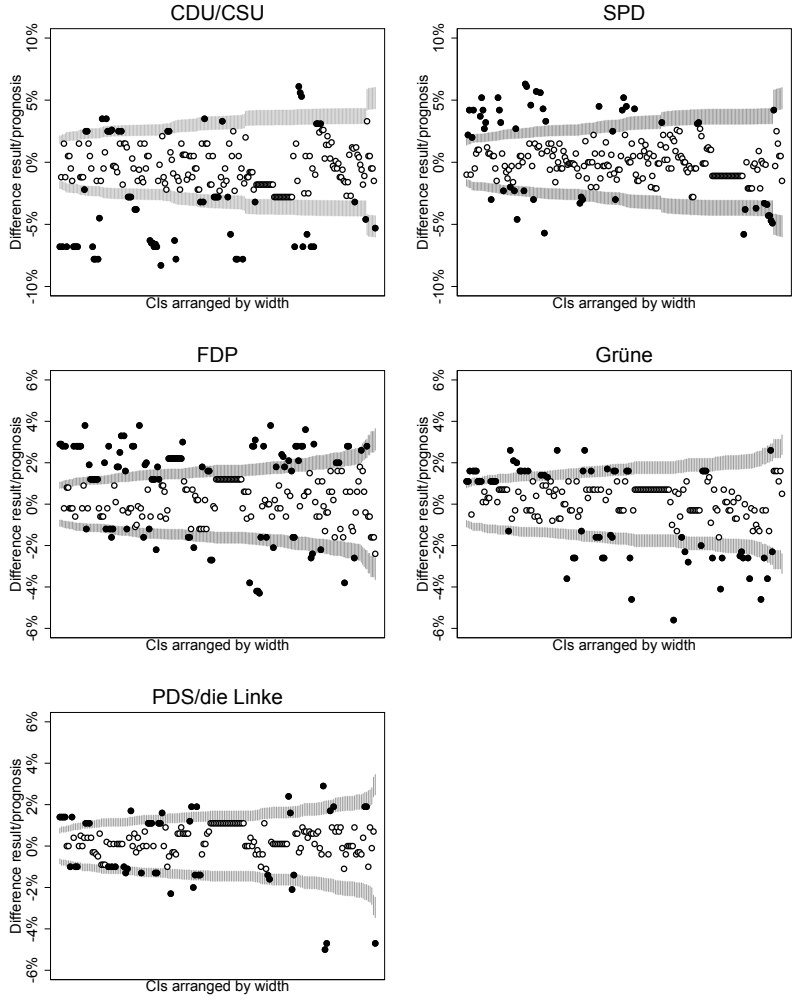
Of central importance for this article is the comparison between the usually applied naïve CIs and the election results. Looking at the coverage of the election results by the calculated naïve CIs, the result is clear (cf. Figure 2): The naïve estimates of the CIs are useless. The aspired confidence level of 95% is missed by far for all parties. Instead of the expected 5%, depending on the party, a minimum of 22% of the CIs do not contain the election result. For the FDP, half of the CIs are affected: instead of 95%-CIs, it comes closer to 50%-CIs (more accurately: 56%-CIs, since 44% of the election results are not contained in the CIs). A coin toss would therefore produce results not much worse than the naïve CIs.

The coverage probability increases greatly when using CIs with design effects. Of those CIs, between 9% and 28% do not contain the election result. These CIs are closer to the usually falsely reported confidence level of 95%, but still far from achieving it.

Figure 3 shows the binomial CIs with and without design effect in comparison for each party. The naïve binomial CIs are distinctly smaller than the corresponding, correctly calculated binomial CIs with design effect.

A consequence of the higher coverage probability is a considerably greater width of the CIs. Figure 4 shows the mean CI widths (CIW) as a dot chart. Half of the correctly computed CIs for the CDU/CSU and SPD have a mean width of more than 7%. FDP, Grüne and Die Linke are roughly at about +/-2%. For most practical applications, this accuracy is not sufficient. If you want to know if a party would

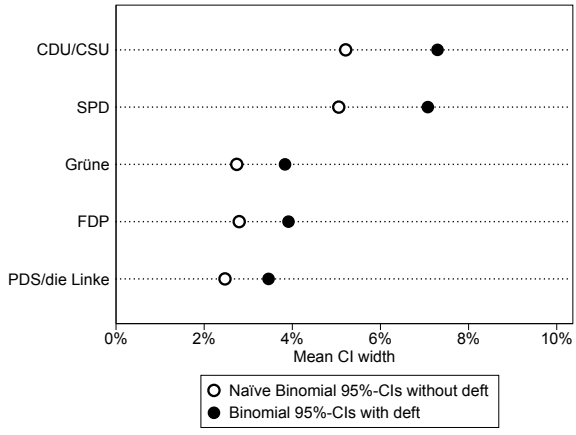
Figure 3: Width of naïve binomial 95%-CIs without deft (inner) in comparison to binomial 95%-CIs with deft (outer) and election results. The naïve CIs without deft corresponding to the results marked as ○ contain the election results; the naïve CIs without deft corresponding to the results marked with ● do not contain the election results.



pass the 5%-electoral threshold, an estimate with a CI from 3% to 7% is factually useless.

This unsatisfactory performance of German pre-election polls becomes more apparent for the number of polls which predicts – given the naïve margin of error – all parties correctly. Statistically, this requires the computation of simultaneous

Figure 4: Dot chart of mean 95%-CI widths



multinomial confidence intervals.[19] For the multinomial CIs, 162 of the 232 Polls (70%) show CIs which all contain the election results. If naïve CIs without deft are used only 67 of the 232 polls (29%) show CIs which all contain the election results. To sum up: Less than a third of the polls would predict all parties within their alleged precision.

The simple fact that small samples, as being used in most polls, cannot deliver the required accuracy for small parties seems to be ignored outside statistics. In general, the width of a CI can be determined given the sample size. If the approximate percentage of votes and the design effect are known, the sample size required for the desired precision can be computed.[20] For a proportion of p=0.4, the width

19 CIs computed for pre-election polls usually assume binomially distributed characteristics. Pre-election polling in Germany has to deal with more than two parties. Therefore, the assumption of binomial distributions is inappropriate, when the results of a pre-election poll are investigated for all parties simultaneously. In this case, it would be appropriate to apply simultaneous multinomial CIs (Ulmer 1989, 1994). Calculating simultaneous multinomial CIs is more demanding than calculating binomial CIs. The easiest approach is the method suggested by Goodman (1965: 250-251). Here, the simultaneous CIs are adjusted according to the number of CIs calculated. For four parties, this would result in a correction factor of 2.498, and 2.576 for five parties. As correction factor, the z-value of $z_{1-\alpha/2}$, as used for a single CI (1.96 for a 95% interval) is replaced by a z-value of $z_{1-\alpha/(2k)}$, where k equals the number of parties. Combined with the assumed design effect of 1.4, the resulting CI for five parties is:

$$\left[p_i - 2.576 * 1.4 * \sqrt{\frac{p_i(1-p_i)}{n}}; p_i + 2.576 * 1.4 * \sqrt{\frac{p_i(1-p_i)}{n}} \right].$$

20 Since the factors ρ, *deft* and $z_{1-\alpha/2}$ are constant, the width of the CI is determined exclusively by $\sqrt{p(1-p)n^{-1}}$.

of a simultaneous CI for n=1000, and a design effect of 1.4 will be 8.5%. To halve the width of the CI, the sample size has to be quadrupled (e.g. Bortz: 2005: 105). Consequently, the width of a CI for 4000 respondents is 4.3%. 16000 respondents provide a CIW of 2.1%, 64000 respondents a width of 1.1%. The width of the CIs is therefore a linear function of the square root of the number of respondents, which transforms the problem of precision to a financial problem. Given the current options of fieldwork in Germany, a sample of 16000 to 64000 respondents cannot be completed within one or two weeks, as required by pre-election polling (Schnell 2012: 385-386).

Even if the resources of all major companies could be pooled, this survey would fail due to the inadequate costs: a pre-election poll of this scale would cost more than €500000.[21] For a still inaccurate estimate, this is not likely to be acceptable to any sponsor.

6 Alternative Explanations for the Failure of Pre-Election Polls

There are two possible alternative explanations for the results of this study. Obviously, opinion changes in the electorate between the end of fieldwork and the election could produce seemingly erroneous results. A less obvious explanation for our result is an increase in accuracy of the pre-election polls during the observed period from 1957 until 2013. The performance of a scientific technique should improve over time. Therefore, worse results would be expected for older polls. Both mechanisms will be examined in more detail.

The literature on pre-election polls sometimes mentions a *last-minute swing* to explain discrepancies between poll results and election results (Roth 2008: 174).[22] Given this hypothesis, a decreasing amount of error would be expected for pre-election polls closer to the election date. This hypothesis is supported by the US results reported by Crespi (1988: 135-136, 166). His results show a significant correlation of r=0.21 between pre-election poll error (difference poll/election result) and the time interval in days before the election. However, although temporal proximity to the election represents the best predictor, his multiple regression model for the difference between polling and election results explains only 12% variance (Crespi 1988: 167). For German data, Groß (2010: 204-212) observes much longer temporal distances of up to one year, and reports a weak, but significant curvilinear

21 For approaches using pooled micro-data of pre-election polls see Park et al. (2004) and Jackman (2005).

22 Occasionally, the mechanism of the "spiral of silence" is mentioned. However, the meta-analysis of all available empirical studies by Glynn et al. (1997) do not give much support for this hypothesis.

Figure 5: Absolute error of the poll results depending on the number of days to the election. The scatterplot-smoother is Loess with f=0.8

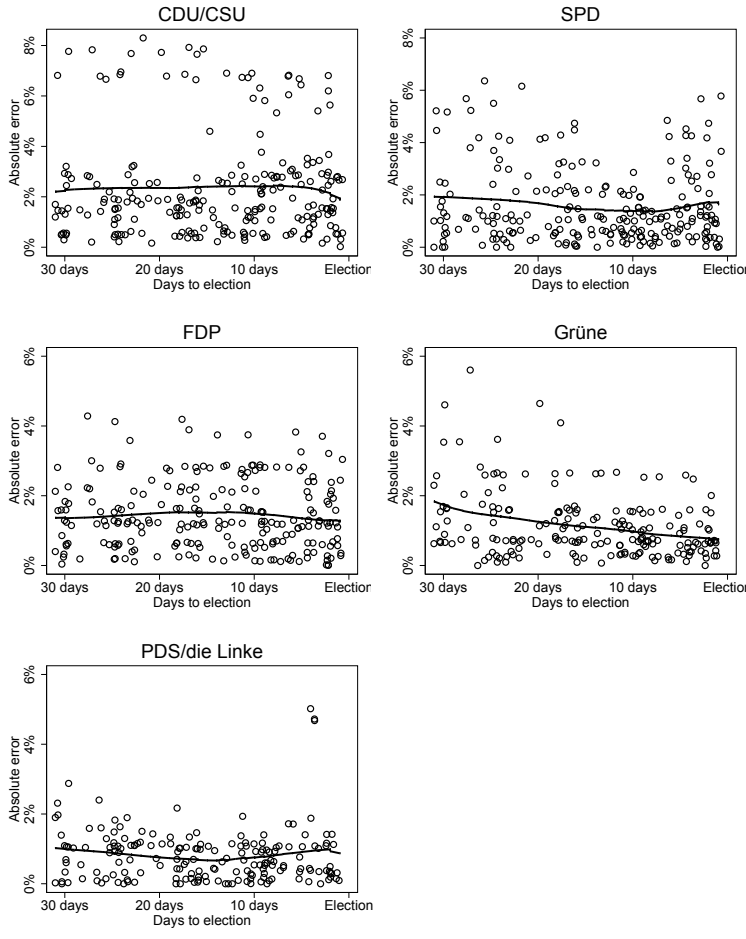
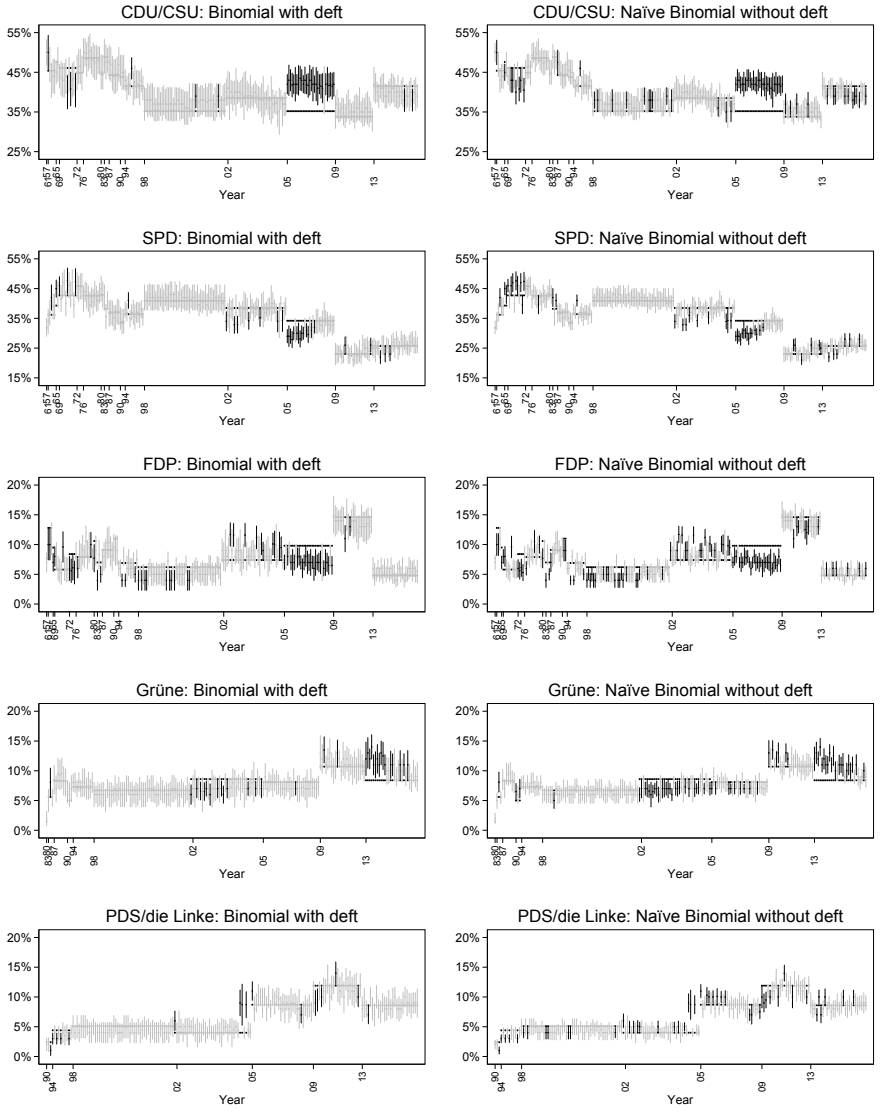


Figure 6: Performance of naïve binomial 95%-CIs without deft (right), in comparison to binomial 95%-CIs with deft (left) over time. Pre-election polls are arranged chronologically. Gray CIs contain the election result, black CIs do not contain the election result. Point estimates are shown as dots.



correlation between temporal distance and error. Our data neither shows a linear nor a nonlinear relationship (cf. Figure 5).[23] Last-minute swings do not seem to be of primary importance for the inaccuracy of the pre-election polls.

The hypothesis of increasing poll accuracy is not supported by the data. This is shown in Figure 6.

7 Conclusions

The comparison of reported margins of error with the actual errors of German pre-election polls between 1957 and 2013 shows disillusioning results: the observed inaccuracy is considerably greater than the published margins of error suggest. The computations of the usual binomial CIs, as taught in most introductory statistical textbooks, is misleading at best. The actual coverage is far below the desired 95%. For some of the small parties, the result is only marginally more accurate than a coin toss. At least for Germany, pre-election polls are not a useful forecasting tool. Applying the statistically more appropriate binomial CIs with design effects, the coverage increases, but at the cost of enlarging the already wide CIs.

Therefore, the results reported here suggest the following conclusions:

- Pre-election polls are not suitable as introductory statistical textbook examples. The formulas to calculate naïve CIs for binomial distributions that are widely used in those textbooks are inappropriate and produce results that are not in accordance with the empirical coverage probabilities.
- The size of the correctly computed CIs (binomial CIs with design effects) make them useless for practical purposes.
- German polling companies rarely report the necessary information for the evaluation of their polling results.

The ad-hoc theoretical weighting of the polling results is neither documented, nor helpful: Although in some cases a reduction of error by theoretical weighting cannot be excluded a priori, systematic evidence favoring theoretical weighting has not been published.

Sampling errors represent only one component of the MSE mentioned in section 2. It is, however, the only component that is quantifiable without a special survey design. Under simplified assumptions, other components may also be esti-

23 A weak effect can only be observed for one of the small parties (Grüne). This effect is due to the election in 2013. Even with these outliers, the temporal proximity explains less than 10% of the variance for this party. The effects remain stable even if not absolute errors, but relative absolute errors are used.

mated, but this would still require more complex designs. The TSE model is therefore used as a regulating idea, rather than an analytical model (Schnell 2012: 388). Assuming that all other components of the TSE do not affect the polling results, the electoral results should be covered by about 95% of the correctly calculated CIs. The data in Figure 2 clearly contradicts this assumption.

The observed low coverage rate of the confidence intervals of German pre-elections could be due to biased estimates, larger variance of the estimators or changing population parameters.[24] Since we eliminated the standard explanation with last minute swings in section 6, biased estimates and increased variances are likely. In our view, the failure of the pre-election polls is primarily due to the limits of measurement of the dependent variable (*Sonntagsfrage*) and the confounding with a second variable of interest, the likelihood of voting. Finally, interviewer effects may be the cause of the increased variance of the estimates (Schnell/Kreuter 2005).

The standard model for pre-election polling in Germany is based on small samples and neither uses a tested theoretical model for coverage errors, nonresponse, electoral participation nor a model for the final decision of undecided voters. Empirically, this model fails far more often than it succeeds.

Acknowledgement

We want to thank Jochen Groß for providing the initial data set, Dipl. Bib. Heidi Dorn for providing the missing sample sizes for 84 studies and the two anonymous reviewers for their very helpful comments.

References

AAPOR. (2009). An evaluation of the methodology of the 2008 pre-election primary polls: A report of the ad hoc committee on the 2008 presidential primary polling. Lenexa: American Association for Public Opinion Research.

AAPOR. (2010). *AAPOR code of professional ethics & practices*. American Association for Public Opinion Research. Retrieved December 16. 2013 from http://www.aapor.org/AM/Template.cfm?Section=AAPOR_Code_of_Ethics&Template=/CM/ContentDisplay.cfm&ContentID=4248

ADM. (1999). *Standards zur Qualitätssicherung in der Markt- und Sozialforschung*. Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute. Retrieved December 16. 2013 from http://www.adm-ev.de/fileadmin/user_upload/PDFS/QUALI.PDF

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*. Hoboken: Wiley.

24 We are thankful to a reviewer for making this point clear.

Bailar, B. (1983). Interpenetrating subsamples. In N. L. Johnson & S. Kotz (Eds.), *Encyclopedia of Statistical Sciences, Vol. 4*. (pp. 197-201) New York: Wiley.

Behnke, J., Baur, N., & Behnke, N. (2006). *Empirische Methoden der Politikwissenschaft*. Paderborn: Schöningh.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. Hoboken: Wiley.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler.* 6. Edition. Heidelberg: Springer.

Bosch, K. (2012). *Statistik für Nichtstatistiker.* 6. Edition. München: Oldenbourg.

Callegaro, M., & Gasperoni, G. (2008). Accuracy of pre-election polls for the 2006 italian parliamentary election: Too close to call. *International Journal of Public Opinion Research*, 20, 148-170.

Converse, P. E., & Traugott, M. W. (1986). Assessing the accuracy of polls and surveys. *Science*, 234, 1094-1098.

Crespi, I. (1988). *Pre-election polling. Sources of accuracy and error.* New York: Russel Sage Foundation.

DeSart, J., & Holbrook, T. (2003). Campaigns, polls, and the states: Assessing the accuracy of statewide presidential trial-heat polls. *Political Research Quarterly*, 56, 431-439.

Devore, J. L., & Berk, K. N. (2012). *Modern mathematical statistics with applications.* 2. Edition. New York: Springer

Durand, C., Blais, A., & LaRochelle, M. (2004). The polls-review: The polls in the 2002 french presidential election: An autopsy. *Public Opinion Quarterly*, 68, 602-622.

Fahrmeir, L., Künstler, R., Pigeot, I., & Tutz, G. (2007). *Statistik - Der Weg zur Datenanalyse*. 6., Revised Edition. Berlin/Heidelberg: Springer.

Frankovic, K. A., Panagopoulos, C., & Shapiro, R. Y. (2009). Opinion and election polls. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of Statistics: Sample Surveys - Design, Methods and Applications, Vol. 29A*. (pp. 566-595). Amsterdam: Elsevier.

Gehring, U. W. & Weins, C. (2009). *Grundkurs Statistik für Politologen und Soziologen*. 5., Revised Edition Wiesbaden: VS-Verlag.

Glynn, C. J., Hayes, A. F., & Shanahan, J. (1997). Perceived support for one's opinions and willingness to speak out - a meta-analysis of survey studies on the "spiral of silence". *Public Opinion Quarterly*, 61, 452-463.

Goodman, L. A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics*, 7, 247-254.

Groß, J. (2010). *Die Prognose von Wahlergebnissen. Ansätze und empirische Leistungsfähigkeit*. Wiesbaden: VS-Verlag.

Groves, R. M. (1989). *Survey errors and survey costs.* New York: Wiley.

Hilmer, R. (2009). Exit polls - genauer geht's nicht. In H. Kaspar, H. Schoen, S. Schumann & J. R. Winkler (Eds.), *Politik - Wissenschaft - Medien. Festschrift für Jürgen W. Falter.* (pp. 257-267). Wiesbaden: VS-Verlag.

ICC/ESOMAR. (2008). *ICC/ESOMAR international code of market and social research*. International Chamber of Commerce/ESOMAR. Retrieved December 16. 2013 from http://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ICCESOMAR_Code_English_.pdf

Jackman, S. (2005). Pooling the polls over an election campaign. *Australian Journal of Political Science,* 40, 499-517.

Keeter, S., Kennedy, C., Dimock, M., Best, J., & Craighill, P. (2006). Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. *Public Opinion Quarterly*, 70 (Special Issue), 759-779.

Keeter, S., Miller, C., Kohut, A., Groves, R. M., & Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64, 125-148.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Klammer, B. (2005). *Empirische Sozialforschung - Eine Einführung für Kommunikationswissenschaftler und Journalisten*. Konstanz: UVK.

Krishnamoorthy K. & Peng, J. (2011). Improved closed-form prediction intervals for binomial and Poisson distributions. *Journal of Statistical Planning and Inference,* 141, 1709–1718.

Lau, R. R. (1994). An analysis of the accuracy of "trial heat" polls during the 1992 presidential election. *Public Opinion Quarterly*, 58, 2–20.

Luderer, B. (2008). *Klausurtraining Mathematik und Statistik Für Wirtschaftswissenschaftler: Aufgaben - Hinweise - Lösungen*, 3., Revise Edition. Wiesbaden: Vieweg+Teubner.

Lusinchi, D. (2012): "President" Landon and the 1936 literary digest poll: Were automobile and telephone owners to blame? *Social Science History*, 36, 23-54.

Lynn, P. & Jowell, R. (1996). How might opinion polls be improved?: The case for probability sampling. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159, 21-28.

Magalhães, P. C. (2005). Pre-election polls in Portugal: Accuracy, bias, and sources of error, 1991-2004. *International Journal of Public Opinion Research*, 17, 399-421.

Mitofsky, W. J., (1998). Was 1996 a worse year for polls than 1948? *Public Opinion Quarterly*, 62, 230-249.

Oestreich, M., & Romberg, O. (2012). *Keine Panik vor Statistik! Erfolg und Spaß im Horrorfach nichttechnischer Studiengänge*. 4., Updated Edition. Wiesbaden: Vieweg+Teubner.

O'Muircheartaigh, C., & Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 161, 63-77.

Park, D. K., Gelman, A., Bafumi, J.(2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, 12, 375-385.

Roth, D. (2008). *Empirische Wahlforschung - Ursprung, Theorien, Instrumente und Methoden*. Wiesbaden: VS-Verlag.

Sanders, D. (2003). Pre-election polling in Britain, 1950-1997. *Electoral Studies*, 22, 1-20.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer.

Schnell, R. & Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21, 389-410.

Schnell, R. (1997). *Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen*. Opladen: Leske+Budrich.

Schnell, R. (2012). *Survey-Interviews. Standardisierte Befragungen in den Sozialwissenschaften*. Wiesbaden: VS-Verlag.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 101-113.

Sison, C. P., & Glaz, J. (1995). Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association*, 90, 366-369.

Ulmer, F. (1989). *Wahlprognosen und Meinungsumfragen und der Ablasshandel mit den Prozentzahlen: der Lotteriecharakter des repräsentativen Querschnittes - Sonderdruck aus Heft 30./31. Jahrgang der Zeitschrift für Markt-, Meinungs- und Zukunftsforschung*. Tübingen: Wickert-Institute/Demokrit-Verlag.

Ulmer, F. (1994). *Der Dreh mit den Prozentzahlen*. Wuppertal: Bergische Universität GH Wuppertal.

United Nations. (2005). *Household sample surveys in developing and transition countries*. New York: United Nations Publication.

Walsh, E., Dolfin, S., & DiNardo, J. (2009). Lies, damn lies, and pre-election polling. *American Economic Review*, 99, 316-322.

Wang, H. (2008). Exact confidence coefficients of simultaneous confidence intervals for multinomial proportions. *Journal of Multivariate Analysis*, 99, 896-911.

Wüst, A. M. (2003). Stimmung, Projektion, Prognose? In A. M. Wüst (Eds.), *Politbarometer.* (pp. 83-107). Wiesbaden: VS-Verlag.

Wüst, A. M. (2010). Exit Poll. In D. Nohlen & R.-O. Schultze (Eds.), *Lexikon der Politikwissenschaft, Band 1 A-M*. 4., Updated and Revised Edition. (pp. 242-243). München: C. H. Beck.