Some Statistics concerning the Austrian Presidential Election 2016

Erich Neuwirth[*] Walter Schachermayer[†]

September 21, 2016

Abstract

The 2016 Austrian presidential runoff election have been repealed by the Austrian constitutional court. The results of the counted votes had yielded a victory of Alexander van der Bellen by a margin of 30.863 votes as compared to the votes for Norbert Hofer. However, the constitutional court found that 77.769 votes were "contaminated" as there have been - at least on a formal level - violations of the legal procedure when counting those votes. For example, the envelopes were opened prematurely, or not all the members of the electoral board were present during the counting etc. Hence the court considered the scenario that the irregular counting of these votes might have caused a reversal of the result as *possible.* The constitutional court sentenced that this *possibility* presents a sufficient irregularity in order to order a repetition of the entire election.

While it is, of course, *possible* that the irregular counting of those 77.769 votes reversed the result, we shall show that the probability, that this indeed has happened, is ridiculously low.

[*] Fakultät für Informatik, Universität Wien, Währinger Straße 29, A-1090 Wien, Austria, erich.neuwirth@univie.ac.at

[†] Fakultät für Mathematik, Universität Wien, Oskar-Morgenstern-Platz 1, A-1090 Wien, Austria, walter.schachermayer@univie.ac.at and the Institute for Theoretical Studies, ETH Zurich. Partially supported by the Austrian Science Fund (FWF) under grant P25815, the Vienna Science and Technology Fund (WWTF) under grant MA09-003 and Dr. Max Rössler, the Walter Haefner Foundation and the ETH Zurich Foundation.

1 Introduction

On May 22, 2016 the Austrians voted in a run-off election between Norbert Hofer (candidate 1) and Alexander van der Bellen (candidate 2). The result after counting the votes was 49.7 : 50.3 in favor of van der Bellen. For the precise data we refer to [1].

The party supporting the candidate Norbert Hofer subsequently appealed to the Austrian constitutional court, claiming irregularities in the procedure of counting the mail votes.

Here are the details. In Austria there are two ways of casting one's vote. Either by showing up personally at the poll site and delivering the vote into the ballot box (ballot voting), or by sending the vote by mail during a well-defined period preceding the voting day (mail voting).

The allegation of Hofer's party was that in some districts the counting of the mail votes violated the procedure stipulated by the law. For example, the letters of the outer envelopes containing these votes (in an inner envelope) should only not be opened before 9:00 a.m. of the subsequent Monday, May 23, the reason being that the electoral board for the mail votes for districts only is called to duty for this time. By opening these letters prematurely these votes became invalid, as argued by the alleging party. Several other accusations were made, involving different degrees of severity [9].

K

The constitutional court carefully investigated these accusations and concluded that in 11 of the 117 voting districts there have indeed happened violations of the law during the procedure of counting the mail votes. The court sentenced that in total there were 77.769 mail votes counted in an irregular way. The central argument of the court in favor of ordering a repetition of the election was that there was the *possibility* that manipulations on such a number of votes might have led to a reversal of the result. After all, the margin was only 30.863 votes.

The constitutional court states explicitly in its finding that there was no evidence that there actually have been manipulations of the votes. What has been proven were several violations of the legally prescribed procedure of counting the votes.

2 The Analysis

Our goal is to obtain a quantitative analysis of the probability that there was in reality a victory of Norbert Hofer which was only turned afterwards – in whatever way – into a victory of Alexander van der Bellen because of wrong-counting the mail votes in the incriminated 11 districts.

To do so, we first compare the results in the $N = 106 = 117 - 11$ “green” or “uncontaminated” districts where the court did not find violations of the legal procedures, with the $M = 11$ “red” or “contaminated” districts districts where the court found violations of these procedures.

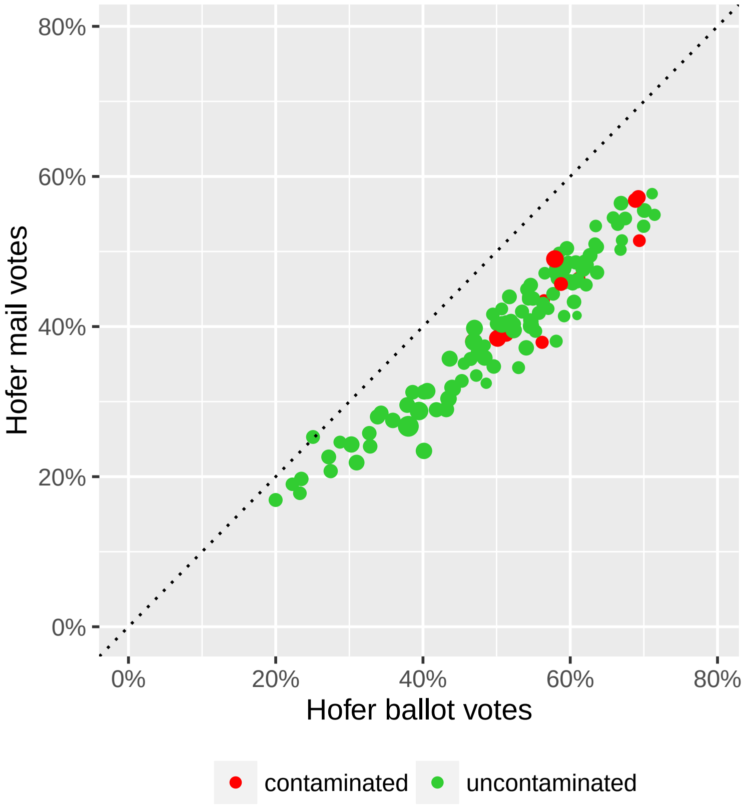

Figure 1: Mail and ballot vote percentages - official results

Each “green” district corresponds to a green dot: on the x-axis we plot the percentage of votes for candidate 1 (Norbert Hofer) among the ballot votes, and on the y-axis the percentage of the votes for candidate 1 among the votes by mail. The picture clearly indicates a linear relation between these two ratios. One also sees that the slope of the regression line is smaller than 1 which does not come as a surprise. Among the voters of Norbert Hofer the propensity to use the possibility of voting by mail is smaller than among the voters of Alexander van der Bellen. We can also observe that the intercept of the regression line essentially vanishes, i.e. the regression line essentially passes through the origin.

While the green dots correspond to the "uncontaminated" districts, the red dots in Figure 1 correspond to the 11 "contaminated" ones. Glancing at Figure 1 one cannot see any alarming behavior of the red dots.

The above Figure 1 corresponds to the *counted* votes. In the subsequent analysis we shall only accept the green dots as valid data. As regards the red dots, we only take their x-coordinate as granted: recall that the x-coordinate corresponds to the percentage of ballot votes in favor of candidate 1. As regards the y-coordinates of the red dots, it is precisely our point to analyze whether the *true* votes gave different results than the *counted* votes in a degree which could have resulted in a reversal of the election result.

As an illustration, Figure 2 below indicates a scenario for the *true* votes which would have yielded a victory for candidate 1 (by a margin of 1 vote). To obtain Figure 2, we have assigned – hypothetically – 15.432 (half the missing 30.863 votes rounded up) proportionally to the 11 "contaminated" districts, and subsequently recalculated the corresponding percentages. This procedure implements a scenario where that many votes have wrongly been counted for candidate 2 instead of candidate 1. For more detailed information we refer to the web site of the first named author [2] (`http://www.wahlanalyse.com/WahlkartenDifferenzenVfGh.html`).

It is evident that a scenario as in diagram 2 does not look very likely to have happened in reality. This diagram only has an illustrative character for our purposes in order to visualize the absurdity of such a scenario. It will not play any role in the subsequent analysis. In particular, we shall not assume a certain given assignment of the missing 30.863 votes to the 11 red districts. We shall only be interested in their *total sum*. Speaking mathematically, we shall eventually calculate the probability distribution of the total number of *true* mail votes in the incriminated districts, conditionally on the given results of the "green" districts, and calculate the probability of the event that they would have resulted in a victory of candidate 1.

To analyze the probability that the *true* votes by mail in the red districts would have yielded a victory for candidate 1, we apply a weighted linear regression model to the green dots in diagram 1.

In fact, since the calculations become easier to write and to program, we use a regression model on the number of votes instead of the percentages, which is mathematically equivalent. Since the expected variation of the votes depends on the total number of votes, the model exhibits heteroskedascity, and we have to use a weighted regression.

Regression models include prediction intervals for observations with known values for the independent variable(s) and known standard deviations (compared to standard deviations of completely known cases). Using these procedures and assuming that the mail vote results in the contaminated districts

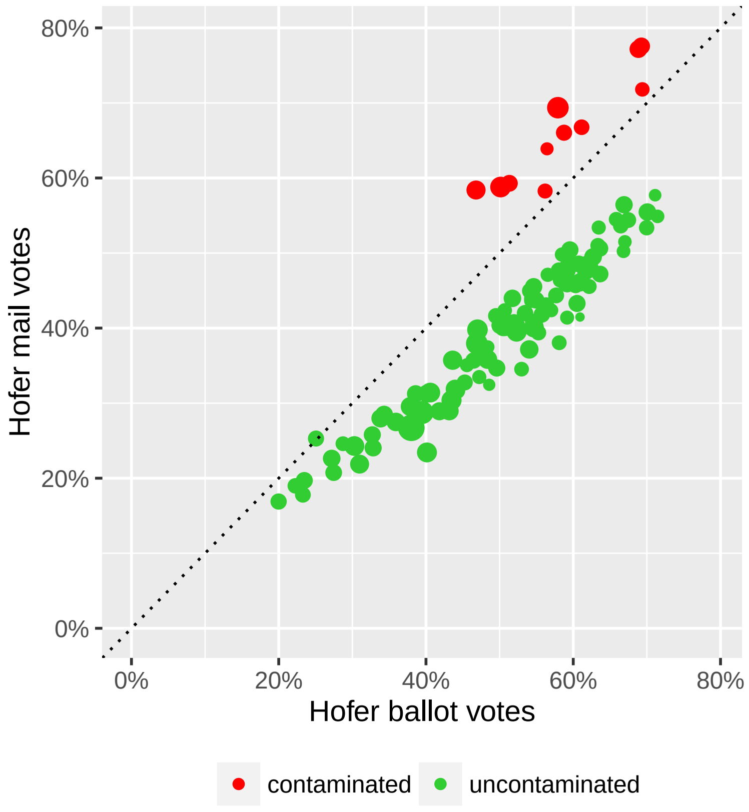

Figure 2: Mail and ballot vote percentages - modified results

follow the model of the uncontaminated districts, we can then compute the distribution of the sum of the expected votes in the contaminated 11 districts (which, under the present model assumptions, follows a rescaled *t*-distribution).

Using the distribution of this expected value, we can then calculate the probability that the *true votes* would have resulted in an election of candidate 1. The numerical value equals $p = 1.322065 \cdot 10^{-10}$.

Actually, it turns out that the sentence of the constitutional court may also lead to a slightly different calculation. Apart from the 11 “contaminated” districts it identified 3 more “dubious” districts where things are not so clear. We refer to [9] for the details. Although the sentence of the court did not take into account these districts, one might argue that – possibly – also in these districts the counting of the mail votes was not reliable. Mathematically speaking, this leads to the consideration of M = 14 “red” and N = 117 -14 = 103 “green” districts. The calculations are identical to the above considered case and lead to a numerical value of $p = 5.151422 \cdot 10^{-8}$.

This modified analysis still gives an extremely low probability which for practical purposes rules out the possibility of manipulations having taken place. It also shows that the results of the analysis are quite robust under slightly different model assumptions.

3 The Model

We build our model following the ideas of model based survey approach, namely we interpret the results of the 106 districts without irregularities as a sample of all mail results and use for prediction of the overall results an estimate for the 11 districts with possible irregularities. This can be seen as as an application of the ratio estimator (see e.g. [8]).

We consider $N = 117 - 11 = 106$ voting districts with a total of t_n valid votes for $n = 1, \ldots, N$. In district n v_n votes were counted for candidate 1 and $\bar{v}_n$ votes for candidate 2 so that $v_n + \bar{v}_n = t_n$.

The votes t_n split into b_n many ballot votes and m_n many mail votes which again are divided into $v_{b,n}(resp.\ v_{m,n})$ many votes for candidate 1 and $\bar{v}_{b,n} = b_n - v_{b,n}$ $(resp.\ \bar{v}_{m,n} = m_n - v_{m,n})$ many votes for candidate 2.

Our objects of interest are the vote numbers

$$v_{b,n} \quad \text{and} \quad v_{m,n}, \quad n = 1, \ldots, N.$$

These numbers denote the counted votes for candidate 1 among the ballot and mail votes respectively. As indicated by the diagrams above, a linear relation between these quantities is justified as model assumption. To make the plots easier to understand, we used percentages instead of votes there.

While the numbers $(v_{b,n})_{n=1}^{N}$ are considered as given data the numbers $(v_{m,n})_{n=1}^{N}$ are considered as realizations of the following random variables:

$$V_{m,n} = k\, v_{b,n} + \epsilon_n, \quad n = 1, \ldots, N \tag{1}$$

Here k is an unknown deterministic number while $(\epsilon_n)_{n=1}^{N}$ are independent centered Gaussian random variables. Variances of votes for parties being proportional to the number of total votes is a standard model assumption in statistical voting analysis procedures (see [3], [5], [6], [7], and [4]).

Assuming independence of vote counts in different districts seems very natural. Independence of the decision of single voters is more difficult to justify. In fact, if this were the case, one could model the sum of votes as a binomial (or multinomial) random variable. We can, however, assume that for small groups of voters (e.g. families) there is a fixed covariance structure of the voting decisions for the members of such a group. Then, the sum of votes for each group of a fixed size has fixed variance. Assuming the the vote sums of different groups are independent, we see that the variance of the vote sums of a district is proportional to the number of groups, and, since we assume the groups to be of essentially equal sizes, also proportional to the number of voters.

Therefore, the variances of our random variables ($V_{m,n}$, have values

$$\mathrm{var}(V_{m,n}) = \mathrm{var}(\epsilon_n) = \sigma^2 m_n \tag{2}$$

for some (unknown) deterministic number $\sigma > 0$.

We are thus facing a heteroskedastic, linear regression model.

Applying standard regression theory, we obtain the estimators $\hat{k}, \hat{\sigma}$ which we consider as random variables. In particular, the estimator $\hat{k}$ follows a (rescaled) t-distribution whose parameters can be explicitly calculated for the given data.

We next consider $M = 11$ many "contaminated" districts, disjoint from the N " uncontaminated" districts.

Assuming that the results $V_{m,j} \quad j = 1, \ldots, M.$ also follow model 1, they can be considered random variables

$$V_{m,j} = k\, v_{b,j} + \epsilon_j, \quad j = 1, \ldots, M.$$

As we do not know the true value of k we have to consider the estimated variable

$$\hat{V}_{m,j} = \hat{k} v_{b,j} + \epsilon_j, \quad j = 1, \ldots, M.$$

The new noise variables $(\epsilon_j)_{j=1}^M$ are such that $((\epsilon_n)_{n=1}^N, (\epsilon_j)_{j=1}^M)$ are independent. The variance of the ϵ_j again is given by $\sigma^2 m_j$.

Finally we consider the sum

$$\hat{V} = \sum_{j=1}^{M} \hat{V}_{m,j},$$

the total ballot result of candidate 1 in the contaminated districts, $v_b = \sum_{j=1}^M v_{b,j}$, the total number of votes there, $m = \sum_{j=1}^M m_j$.

$\hat{V}$ is the random variable modeling the total mail votes for candidate 1 in the M "contaminated" districts.

The random variable $\hat{V}$ follows the model

$$\hat{V} = \hat{k}\, v_b + \epsilon$$

where ϵ is a centered Gaussian variable with variance $\sigma^2 m$.

A standard tool of regression theory allow us to compute a prediction interval for $\hat{V}$.

If σ^2 were known, $\hat{V} = \hat{k}v_b + \epsilon$ were distributed with mean $\hat{k}v$ and variance of $\hat{V}$ equal to $\sigma^2\left(\frac{v_b^2}{\sum_{n=1}^N \frac{v_{b,n}^2}{m_n}} + m\right)$

Using this fact we use the regression model estimate for $\hat{\sigma}$ as substitute for the unknown constant σ. Using this, the random variable

$$\frac{\hat{V} - \hat{k}v_b}{\hat{\sigma}\sqrt{\frac{v_b^2}{\sum_{n=1}^N \frac{v_{b,n}^2}{m_n}} + m}}$$

follows a t-distribution with 105 degrees of freedom.

We compare this random variable with the critical number $\tilde{V}$ which would be necessary for candidate 1 reversing the result. Finally we compute

$$\mathbb{P}[V \geq \tilde{V}] \tag{3}$$

which can be computed explicitly.

4 Confidence intervals for prediction

We use results for the standard heteroskedastic linear model

$$y = X\beta + \epsilon$$

with covariance matrix $(cov)(\epsilon) = \sigma^2 W$.

In this model, the best linear unbiased estimator for β is

$$\hat{\beta} = (X'W^{-1}X)^{-1}X'W^{-1}y$$

The covariance of this estimator is

$$\begin{aligned}\mathrm{cov}(\hat{\beta}) &= (X'W^{-1}X)^{-1}X'W^{-1}\mathrm{cov}(y)((X'W^{-1}X)^{-1}X'W^{-1})' \\ &= (X'W^{-1}X)^{-1}X'W^{-1}\sigma^2 WW^{-1}X(X'W^{-1}X)^{-1} \\ &= \sigma^2(X'W^{-1}X)^{-1}\end{aligned}$$

In our case, X is the Nx1-matrix $(v_{b,n})_{n=1}^N$ and W is the diagonal matrix $\mathrm{diag}((m_n)_{n=1}^N)$. The parameter β in our case is the scalar k and its estimator is $\hat{k}$, and $\mathrm{cov}(\hat{\beta})$ becomes $\mathrm{var}(\hat{k})$

Therefore $X'W^{-1}X = \sum_{n=1}^{N} v_{b,n}\frac{1}{m_n}v_{b,n} = \sum_{n=1}^{N}\frac{v_{b,n}^2}{m_n}$ and

$$\mathrm{var}(\hat{k}) = \sigma^2\frac{1}{\sum_{n=1}^{N}\frac{v_{b,n}^2}{m_n}}$$

We want to compute the distribution of $\hat{V} = \hat{k}v_b + \epsilon$ with $\mathrm{E}(\epsilon) = 0$ and $\mathrm{var}(\epsilon) = \sigma^2 m$.

We have

$$\begin{aligned}\mathrm{E}(\hat{k}v_b + \epsilon) &= \mathrm{E}(\hat{k}v_b) + \mathrm{E}(\epsilon) = kv_b \\ \mathrm{var}(\hat{k}v_b + \epsilon) &= \mathrm{var}(\hat{k}v_b) + \mathrm{var}(\epsilon) = \sigma^2\left(\frac{v_b^2}{\sum_{n=1}^{N}\frac{v_{b,n}^2}{m_n}} + m\right)\end{aligned}$$

Since σ^2 is unknown, we have to replace it by the estimator $\hat{\sigma}^2$ and then $\frac{\hat{V}-\hat{k}v_b}{\hat{\sigma}\sqrt{\frac{v_b^2}{\sum_{n=1}^{N}\frac{v_{b,n}^2}{m_n}}+m}}$ follows a t-distribution with $N-1$ degrees of freedom, an from that confidence intervals for $\hat{V}$ are easily derived.

5 The Results

All the data and the code for performing our analysis can be found at `https://github.com/neuwirthe/AustrianPresidentialElection`.

Hofer had 34479 votes in the contaminated districts, and he would need additional 15432 votes, so that in total he needs $\tilde{V} = 34479 + 15432 = 49911$ votes to overturn the result.

Using the R code from the URL above to compute the probability of a result overturning the result in favor of Hofer, we get the value

$$1.322065 \cdot 10^{-10}.$$

References

[1] http://www.bmi.gv.at/cms/BMI_wahlen/bundespraes/bpw_2016/

[2] http://www.wahlanalyse.com/WahlkartenDifferenzenVfGh.html

[3] Bruckmann, G. (1966). *Schätzung von Wahlresultaten aus Teilergebnissen* Wien: Physica-Verlag.

[4] Ledl, Th. (2007). *Modellierung von Wechselwählerverhalten als Multinomialexperiment* Dissertation, Fakultät für Wirtschaftswissenschaften und Informatik, Universität Wien. `http://homepage.univie.ac.at/thomas.ledl/download/Dissertation.pdf`

[5] Neuwirth, E. (1984). *Schätzung von Wählerübergangswahrscheinlichkeiten* In: M. Holler (Hg.) Wahlanalyse – Hypothesen, Methoden und Ergebnisse. München: tuduv-Buch.

[6] Neuwirth, E. (1994). *Prognoserechnung am Beispiel der Wahlhochrechnung* In: P. Mertens. (Hg.) Prognoserechnung. physica-Verlag, Würzburg-Wien.

[7] Neuwirth, E. (2012), *Wahlhochrechnung: ein kurzer Überblick über den Einsatz bei bundesweiten Wahlen in Österreich*, in: Österreich 2032 (Festschrift zum 80. Geburtstag von Gerhart Bruckmann), Hrsg: Lutz, W. und Strasser, H., Verlag der österreichischen Akademie der Wissenschaften, Wien 2012.

[8] Valliant, R., Dorfman, A. H. and Royall, R. M. (2000). *Finite Population Sampling and Inference*, Wiley,

[9] Verfassungsgerichtshof, Freyung 8, A-1010 Wien, *W I 6/2016-125*,
1. Juli 2016,
`https://www.vfgh.gv.at/cms/vfgh-site/attachments/5/7/8/CH0003/CMS1468412977051/w_i_6_2016.pdf`