

On the verification of a large petitionnement

H. W. J. Blöte and J. M. J. van Leeuwen

October 22, 1998

Abstract

The problem of checking the integrity of a large petition is analysed. It is argued that this is a very involved procedure

1 Introduction

Frequently society is confronted with an issue which generates strong opinions pro and con. The segment that feels that their voice is not sufficiently heard, often has recourse to the submission of a petition to the government. The force of the petition is then measured by the number of endorsing signatures. The number of signatures can be of the order of 10^6 . The opposing fraction then sometimes argues that the acquisition of signatures is fraudulent e. g. that overzealous people have submitted several times their signature in order to enlarge the number. The problem is how large a sample must be in order to check reliably whether such malpractices have taken place.

2 Intuitive Calculation

Suppose we have $2 * 10^6$ signatures, of which half are bona fide and half not. I. e. the mala fide signatures consist of $0.5 * 10^6$ equal pairs. Now we pick out of the set a sample of 2000 signatures and check them on doubles. With a large probability one will select 1000 signatures which are single and another 1000 which are a member of a pair. The question is how large is the probability that a pair is found in this set of 1000 malafide signatures? For a specific pair to occur the probability is 10^{-6} i. e. the product of 10^{-3} for each member of the pair. The probability that this pair will not occur in the sample is therefore $1 - 10^{-6}$ and since there are $0.5 * 10^6$ malafide pairs the probability that none of them will show up is

$$(1 - 10^{-6})^{0.5 * 10^6} \simeq e^{-0.5}$$

One may generalize this result to the probability of finding precisely k pairs in the sample. By the same reasoning the probability of finding a specific set of k pairs and none of the other $0.5 * 10^6 - k$ equals

$$10^{-6k} e^{-0.5}$$

This has to be multiplied by the number of ways that k pairs can be selected out of the $0.5 * 10^6$ pairs. Thus the probability p_k on finding k pairs is

$$p_k = \binom{0.5 * 10^6}{k} 10^{-6k} e^{-0.5} \simeq \frac{1}{k!} 2^{-k} e^{-0.5}$$

The results of this calculation are summarized in the table below.

signatures		probability
double	single	
0	2000	60.7%
1	1998	30.3%
2	1996	7.6%
3	1994	1.3%
all other cases		0.1%

A further generalization results by considering N signatures which come from $M < N$ supporters who produce $2M - N$ single signatures and $N - M$ doubles. We take a sample of n signatures which (for $1 \ll n \ll N, M$) contains

$$\frac{n}{N}(2M - n) \text{ bona fide} \quad \text{and} \quad \frac{2n}{N}(N - M) \text{ mala fide}$$

signatures. We thus have a set of $N - M$ pairs of mala fide signatures out of which we have selected $2(N - M)/N$ dubious signatures. The probability that both members of a specific pair are drawn is $(n/N)^2$ and so the probability that none of the pairs are drawn is

$$\left[1 - \left(\frac{n}{N}\right)^2\right]^{N-M} \simeq e^{-\frac{n^2(N-M)}{N^2}}.$$

Likewise the probability on a set of k pairs and none of the other pairs is

$$p_k = \frac{1}{k!} \left[\frac{n^2(N-M)}{N^2}\right]^k e^{-\frac{n^2(N-M)}{N^2}}. \quad (1)$$

Of course one retrieves the previous expression by inserting the values $N = (4/3)M = 2 * 10^6$ and $n = 2000$. The general distribution is Poissonian with the mean $n^2(N - M)/N^2$.

3 Exact Calculation

The exact calculation considers the number of ways one can draw, from a set of $N - M$ doubles, precisely k doubles and p singles from the remaining $N - M - k$ doubles and the other $n - 2k - p$ of the sample from the set of $2M - N$ singles. All these individual possibilities are binomial coefficients and dividing by the total number of sample n from a set N one arrives at the probability distribution for the pairs

$$p_k = \binom{N}{n}^{-1} \sum_p \binom{N-M}{k} \binom{N-M-k}{p} 2^p \binom{2M-N}{n-2k-p}.$$

The sum over p can be eliminated by introducing a contour integration such that

$$p_k = \binom{N}{n}^{-1} \binom{N-M}{k} \frac{1}{2\pi i} \oint \frac{dz}{z^{n-2k+1}} (1+z)^{2M-N} (1+2z)^{N-M-k}.$$

expression by using the binomial formula for the powers under the integral. The contour integral picks up all the possible combinations labeled by the index p . Since we assume that $1 \ll n \ll M, N$ we may use the saddle points method which leads directly to the preceding intuitive results. For a number of operations the exact expression is easier represented by the generating function

$$P(y) = \sum_k p_k y^k.$$

which is given by

$$P(y) = \binom{N}{n}^{-1} \frac{1}{2\pi i} \oint \frac{dz}{z^{n+1}} (1+z)^{2M-N} (1+2z+yz^2)^{N-M}$$

For instance the proof that the distribution is normalized follows from substituting $y = 1$

$$\sum_k p_k = P(1) = \binom{N}{n}^{-1} \frac{1}{2\pi i} \oint \frac{dz}{z^{n+1}} (1+z)^N = 1.$$

In the same way we find the mean of the distribution as

$$\sum_k p_k k = \left[\frac{dP(y)}{dy} \right]_{y=1} = \frac{n(n-1)(N-M)}{N(N-1)}$$

and the higher moments by higher derivatives at $y = 1$. In the case $1 \ll n \ll M, N$ these results all reduce to the previously found expressions.

4 Conclusion

Important parameters are the falsification factor $x = (N - M)/N$ which is the fraction of irrelevant signatures and the sample size which we express as $n = \alpha\sqrt{N}$. In terms of x and α the main result (1) for the distribution of doubles reads

$$p_k = \frac{1}{k!} (\alpha^2 x)^k e^{-\alpha^2 x} \tag{2}$$

As small numbers are easily attributed to incidental errors one must make the mean $\alpha^2 x$ of reasonable size, say 10. The values in the table for a small sample are very suggestive. Thus a reasonable size is

$$\alpha = \sqrt{10/x}$$

Here one sees the dilemma. Even in the case of a large false fraction, say 50%, one must raise α to 4.5. For a million signatures this means a sample of 4500 which require 10 million pairs to be inspected. For a relative honest petition with a false fraction of 10% this number goes up to 50 million. Moreover enlarging the number of signatures by producing twice a real and controlable one, does not seem a very clever falsification.

We have only calculated the probabilities of finding frauds in the case that signatures occur at most twice. For multiple signers the conclusions are similar, but more difficult to calculate precisely.

The validity of a large petitionnement is presumably better based on the fact that it is difficult to keep a large scale fraud secrete than on an actual verification of the submitted signatures.