THE VALIDATION OF A PETITION BY A SAMPLE SURVEY

J. J. Shuster, University of Florida

ABSTRACT

A student petition needs K signatures to be approved. A total of N signatures are obtained. This paper deals with the question: Are there more than K distinct signatures among the N? Because of time and cost factors, we shall base our inference on the number of invalid signatures in a random sample of n signatures.

1. INTRODUCTION

Consider a petition that requires K distinct signatures for validation. Suppose that N_1 signatures appear once, N_2 signatures appear twice, and so on. The number of distinct signatures, D, is:

$$D = \sum_{i=1}^{\infty} N_i \ , \qquad (1)$$

while the total number of signatures, N, is

$$N = \sum_{i=1}^{\infty} iN_i \ . \qquad (2)$$

Based on a random sample of n names, we wish to test: H_0 : D≤K against H_A: D>K, at the α×100% significance level.

The burden of "proof" is assumed to fall on the people whose interest is to make the petition pass.

As a test statistic, we shall use Y, the total number of signatures, known to be invalid, that fall in the sample. If n_i is the number of signatures appearing i times in the sample,

$$Y = \sum_{i=1}^{\infty} (i-1)n_i \ . \qquad (3)$$

We shall reject H_0 if Y falls below a specified critical value.

For any given D, the distribution on the N_i which stochastically minimizes Y can be shown to be:

$$N_1 = 2D-N \quad N_2 = N-D \quad N_3=N_4=\ldots=0, \qquad (4)$$

where D and N are given by (1) and (2).

The above is obtained by comparing iteratively Y for an arbitrary N_i scheme with Y for the scheme modified by replacing one signature of multiplicity three or more by a signature that duplicates a signature that appears once.

Since we reject for small values of Y, the significance level for the distribution of N_i given in (4), is at least as large as that of any other distribution.

We now restrict our attention to the distribution given in (4). The problem may be stated as follows: We wish to test: $H_0 : N_2 = R = N-K$ against $H_A: N_2<R$. The test statistic is

$$Y = \text{number of duplicated names in sample.} \qquad (5)$$

Des Raj (1961) studied this problem in connection with matching lists and obtained the mean and variance of Y. We shall give an expression for the exact distribution of Y, and derive a Poisson approximation. Barton (1958) has obtained Poisson limiting distributions in other "matching problems".

2. THE DISTRIBUTIONAL RESULTS

In this section we assume that (4) holds.

Let X be the number of signatures in the sample that come from the $2N_2$ duplicated signatures. By elementary counting rules, we obtain under H_0 :

$$P[Y=y|X=x] = \binom{R}{y}\binom{R-y}{x-2y}2^{x-2y}\left[\binom{2R}{x}\right]^{-1}, \qquad (6)$$

and hence, under H_0 :

$$P[Y=y] = \sum_{x=2y}^{\infty}\binom{R}{y}\binom{R-y}{x-2y}\binom{N-2R}{n-x}2^{x-2y}\left[\binom{N}{n}\right]^{-1}, \qquad (7)$$

where $\binom{a}{b}=0$ unless a,b are non-negative integers with b≤a.

Since N,n, and R will tend to be large, the exact probability given in (7) is difficult to compute directly.

The following theorem is therefore useful:

Theorem: Let N,n, and R → ∞ in such a way that for appropriate constants λ and p:

$$(a) \quad N = n^2/\lambda \ , \qquad (8)$$

$$(b) \quad R = Np = n^2p/\lambda \ . \qquad (9)$$

Then under H_0 :

$$P[Y=y] \to (\lambda p)^y e^{-\lambda p}/(y!), \qquad (10)$$

the Poisson distribution with mean λp.

Proof: Let

$$t = [2np]. \qquad (11)$$

By Stirlings approximation (see Feller (1957), page 54), and simplification, we obtain from (6)

$$P[Y=y|X=t] \sim (\lambda p)^y e^{-y} A_n B_n C_n D_n/(y!), \qquad (12)$$

where

$$A_n = \left(1 - \frac{\lambda}{n}\right)^{-t} \to e^{2\lambda p}, \tag{13}$$

$$B_n = \left(1 - \frac{2\lambda}{n} + \frac{\lambda y}{pn^2}\right)^{t} \to e^{-4\lambda p}, \tag{14}$$

$$C_n = \left(1 - \frac{y}{pn}\right)^{-t} \to e^{2y}, \tag{15}$$

and

$$D_n = A_n^{-2R/t} B_n^{-R/t} \to e^{\lambda p - y}. \tag{16}$$

Hence,

$$P[Y=y|X=t] \to (\lambda p)^y e^{-\lambda p}/(y!). \tag{17}$$

Next, from (6), we obtain for positive integers x and m:

$$\frac{P[Y=y|X=x+m]}{P[Y=y|X=x]} =$$

$$\prod_{j=0}^{m-1} \frac{(x+m-j)(R-x+y-j)}{(x+m-2y-j)(R-\frac{1}{2}(x+j))}. \tag{18}$$

Let $\delta>0$ be an arbitrary positive number and let

$$U(\delta,t) = \left[\frac{(t-\delta t^{\frac{1}{2}})(R-t+y)}{(t-\delta t^{\frac{1}{2}}-2y)(R-\frac{1}{2}(t+\delta t^{\frac{1}{2}}))}\right]^{\delta t^{\frac{1}{2}}}, \tag{19}$$

$$L(\delta,t) = \left[\frac{R-t+y-\delta t^{\frac{1}{2}}}{R-\frac{1}{2}t}\right]^{\delta t^{\frac{1}{2}}}, \tag{20}$$

and

$$M(\delta,t) = \max\{U(\delta,t), [L(\delta,t)]^{-1}\}. \tag{21}$$

Then by elementary analysis,

$$[M(\delta,t)]^{-1} \le \frac{P[Y=y|X=x]}{P[Y=y|X=t]} \le M(\delta,t), \tag{22}$$

for every integer x such that

$$|x-t| \le \delta t^{\frac{1}{2}}, \tag{23}$$

where t is given in (11).

Under the conditions of the limiting process, it is readily seen that

$$M(\delta,t) \to 1. \tag{24}$$

Let $\epsilon>0$ be arbitrary, and choose δ such that the hypergeometric random variables X satisfy (for all t given by (11)):

$$P[t-\delta t^{\frac{1}{2}} < X < t+\delta t^{\frac{1}{2}}] > 1-\epsilon. \tag{25}$$

This can be done since for all t,

$$|E(X)-t| < 1 \text{ and } Var(X) < t+1. \tag{26}$$

The value of δ may be obtained from (26) and Chebyshev's inequality.

Let A be the event defined within the probability statement (25).

By (17), (22), (23) and (24),

$$P[Y=y|X\epsilon A] \to (\lambda p)^y e^{-\lambda p}/(y!). \tag{27}$$

However,

$$(1-\epsilon)P[Y=y|X\epsilon A] \le P[Y=y] \le P[Y=y|X\epsilon A]+\epsilon. \tag{28}$$

Since ϵ is arbitrary, the Poisson limiting distribution of Y is immediate from (27) and (28).

3. FIXED FRAME ANALYSIS

Suppose that of the N signatures that have been collected, R are invalid because of duplication of valid signatures, while Q are invalid because they are "illegal" signatures. The illegal signatures in the sample can be spotted with probability one, assuming that a frame is available. (For example, a list of students, homeowners, etc. can be used to check each signature in the sample.)

If we let Z be the number of signatures within the sample and off the frame, and Y be as above, the number of frame signatures in sample, known to be invalid, we estimate T, the total number of invalid signatures by

$$\hat{T} = \frac{N(N-1)Y}{n(n-1)} + \frac{NZ}{n}. \tag{29}$$

At its stochastic minimum,

$$E(\hat{T}) = T. \tag{30}$$

$$Var(\hat{T}) =$$

$$\frac{N(N-1)R}{n(n-1)}\left\{1+\frac{(n-2)(n-3)(R-1)}{(N-2)(N-3)} - \frac{n(n-1)R}{N(N-1)}\right\} + \frac{(N-n)Q(N-Q)}{(N-1)n} - \frac{4QR(N-n)}{n(N-2)}. \tag{31}$$

To estimate $Var(\hat{T})$, we replace R by $\frac{N(N-1)Y}{n(n-1)}$ and Q by $\frac{NZ}{n}$ in (31).

4. NUMERICAL EXAMPLES

1. To illustrate the Poisson approximation:

A petition needing K=12,000 names for validation has N=14,115 signatures. Find the rejection region for type I error no larger than .20, for a sample of n=1000 names.

Solution: Our H_0 is: N_2=2115. Hence,

$$\lambda p = 10.6. \tag{32}$$

We reject the hypothesis that $N_2=2115$ in favor of the hypothesis that $N_2<2115$ if the sample reveals fewer than eight invalid signatures.

If in fact $N_2=1000$ and $N_1=12,115$, the null hypothesis would be rejected with probability .867.

One may point out that our testing method is conservative when a name appears three or more times in the sample. Although this is true, such events will be rare in practice. For example, if N=14115, with $N_1 + 2N_2 = 14095$, and $N_{10} = 2$, the probability that a name appears at least three times in the sample of n=1000, is .05.

2. To illustrate the fixed frame analysis, suppose

$$N = 16,000 \quad n = 1,000 \quad Y = 4 \quad Z = 28$$

$\hat{T} = 1473$ (estimated number of invalid signatures)

$\hat{V}(\hat{T}) = 262,456$ (estimated variance of $\hat{T}$)

$\hat{\sigma}(\hat{T}) = 512$ (estimated standard deviation of $\hat{T}$).

ACKNOWLEDGEMENT

The author wishes to thank Mr. Brian Donerly, College of Journalism, for suggesting the problem.

REFERENCES

Barton, D.E., (1958). The matching distribution, Poisson limiting forms and derived methods of approximation. J.Ro.Statist.Soc., Series B 20, 73-92.

Des Raj, (1961). On matching lists by samples, J.Am.Statist.Assoc., 56, 151-155.

Feller, W., (1968). An Introduction to Probability Theory and its Applications, Vol. 1, 3rd. ed. New York: Wiley and Sons.