ON A PROBLEM OF ESTIMATION OCCURRING IN PUBLIC OPINION POLLS

By Henry B. Mann

Ohio State University

To arrive at an estimate of the number of electoral votes that will be cast for a presidential candidate a poll is taken of $\lambda_i N$ interviews in the ith state ($i = 1, \cdots, 48$) where the λ_i are fixed constants > 0 such that $\Sigma\lambda_i = 1$ and the respondent is asked for which candidate he intends to cast his vote. To estimate the number of electoral votes which candidate A will receive, the electoral votes of all the states in which the poll shows a majority for candidate A are added and their sum is used as an estimate for the number of electoral votes which candidate A will receive. In this paper certain properties of this estimate will be discussed. It will be shown that it is a biased but consistent estimate and an upper bound for the bias will be derived. Finally we shall derive that distribution of interviews which minimizes the variance of our estimate.

In all that follows we shall consider the poll as a random or stratified random sample and shall disregard the bias introduced by inaccurate answers. Our results however remain valid as long as the sampling variance is proportional to $\frac{1}{\sqrt{N}}$.

We shall use the following notation:

π_i = proportion of voters in the ith state who intend to vote for candidate A.

$$\epsilon_i = 1 \quad \text{if } \pi_i > \tfrac{1}{2}$$
$$\phantom{\epsilon_i = {}} 0 \quad \text{if } \pi_i < \tfrac{1}{2}$$

w_i = number of electoral votes of the ith state.

p_i, e_i = sample values of π_i and ϵ_i resp.

We shall further exclude the case $\pi_i = \frac{1}{2}$.

The number of electoral votes for candidate A is then given by

$$\sum_{i=1}^{i=48} \epsilon_i w_i = \Gamma.$$

As an estimate of Γ we use the quantity

$$\sum_{i=1}^{i=48} e_i w_i = G. \tag{1}$$

Let ρ_i be the probability that $p_i > \frac{1}{2}$ and hence $e_i = 1$. Let $\lambda_i N = N_i$ be the number of interviews in the ith state. If N_i is not too small then ρ_i is given by

$$\rho_i = \int_{\frac{1}{2}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x-\pi_i)^2/2\sigma_i^2}\, dx. \tag{2}$$

In this formula $\sigma_i = \sqrt{\frac{\pi_i(1-\pi_i)}{N_i}}$ if the sample is an unstratified random sample and may be somewhat less if the sample is a stratified random sample.[1] For our purposes it is sufficient to assume that σ_i is proportional to $\frac{1}{\sqrt{N_i}}$.

We then have $E(e_i) = \rho_i$ and

$$E(G) = E(\textstyle\sum_{i=1}^{i=48} e_i w_i) = \sum_{i=1}^{i=48} \rho_i w_i. \tag{3}$$

Hence G is a biased estimate of Γ. On the other hand[2] $\operatorname*{plim}_{N\to\infty} p_i = \pi_i$ and hence $\operatorname*{plim}_{N\to\infty} e_i = \epsilon_i$ and therefore $\operatorname*{plim}_{N\to\infty} G = \Gamma$. That is to say G is a consistent estimate of Γ.

According to (3) the bias is given by

$$B(N) = \textstyle\sum_{i=1}^{i=48} \epsilon_i w_i - \sum_{i=1}^{i=48} \rho_i w_i = \sum_{i=1}^{i=48} (\epsilon_i - \rho_i) w_i. \tag{4}$$

We have

$$\epsilon_i - \rho_i = -\frac{1}{\sqrt{2\pi}} \int_{(\frac{1}{2}-\pi_i)/\sigma_i}^{\infty} e^{-\frac{1}{2}x^2}\,dx \quad \text{if} \quad \pi_i < \tfrac{1}{2}$$

$$\epsilon_i - \rho_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(\frac{1}{2}-\pi_i)/\sigma_i} e^{-\frac{1}{2}x^2}\,dx \quad \text{if} \quad \pi_i > \tfrac{1}{2}.$$

For a stratified as well as for an unstratified sample σ_i is proportional to $\frac{1}{\sqrt{N_i}}$ and we therefore put

$$\frac{\frac{1}{2} - \pi_i}{\sigma_i} = \begin{cases} \gamma_i\sqrt{N_i} & \text{if} \quad \pi_i < \frac{1}{2} \\ -\gamma_i\sqrt{N_i} & \text{if} \quad \pi_i > \frac{1}{2} \end{cases}. \tag{5}$$

Then we have in both cases

$$|\epsilon_i - \rho_i| = \frac{1}{\sqrt{2\pi}} \int_{\gamma_i\sqrt{N_i}}^{\infty} e^{-\frac{1}{2}x^2}\,dx. \tag{6}$$

We have for $a > 0$

$$\begin{aligned} \int_a^{\infty} e^{-\frac{1}{2}x^2}\,dx &\le h(e^{-\frac{1}{2}a^2} + e^{-\frac{1}{2}(a+h)^2} + e^{-\frac{1}{2}(a+2h)^2} + \cdots) \\ &< e^{-\frac{1}{2}a^2} h(1 + e^{-ah} + e^{-2ah} + \cdots) \\ &= e^{-\frac{1}{2}a^2} \cdot \frac{h}{1 - e^{-ah}} \end{aligned}$$

for every value h.

Since $\lim_{h\to 0} \frac{h}{1 - e^{-ah}} = \frac{1}{a}$ we have

$$\int_a^{\infty} e^{-\frac{1}{2}x^2}\,dx \le \frac{e^{-\frac{1}{2}a^2}}{a} \quad \text{for every } a > 0. \tag{7}$$

---

[1] The variance in public opinion polls is somewhat larger than the random sampling variance due to the fact that a cluster sample is used and not a random sample. For the same reason the estimate p_i of π_i may be biased.

[2] For the notation used here see: H. B. MANN AND A. WALD, "On stochastic limit and order relationships". *Annals of Math. Stat.*, (1943), pp. 217–227.

From (6) and (7) we obtain

$$|\epsilon_i - \rho_i| \leq \frac{e^{-\frac{1}{2}\gamma_i^2 N_i}}{\sqrt{2\pi N_i}\,\gamma_i}. \tag{8}$$

From (4) and (8) we have

$$|B(N)| \leq \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{i=48} w_i \frac{e^{-\frac{1}{2}\gamma_i^2 N_i}}{\gamma_i \sqrt{N_i}}. \tag{9}$$

Formula (9) is valid whenever $\pi_i \neq \frac{1}{2}$ and shows that $B(N)$ converges rapidly to 0 for all values $\pi_i \neq \frac{1}{2}$.

To obtain an approximate idea of the magnitude of the bias we may in (4) replace ϵ_i and ρ_i by their sample values e_i and r_i. The quantity $\sum_{i=1}^{i=48} w_i (e_i - r_i)$ can, however, not be regarded as an estimate of $B(N)$.

We now proceed to compute the standard error of G. We may consider the poll as 48 single experiments where the probability of success in the ith experiment is given by ρ_i where

$$\frac{1}{\sqrt{2\pi}} \int_{\gamma_i\sqrt{N_i}}^{\infty} e^{-\frac{1}{2}x^2}\, dx = \begin{cases} \rho_i & \text{if} \quad \pi_i < \frac{1}{2} \\ 1 - \rho_i & \text{if} \quad \pi_i > \frac{1}{2} \end{cases}.$$

Hence the variance of G is given by

$$\sigma^2 = \sum_{i=1}^{i=48} \rho_i (1 - \rho_i) w_i^2. \tag{10}$$

As an estimate of σ^2 we can use the quantity S^2 obtained by replacing ρ_i by its sample value.

We shall consider that distribution of interviews as best which minimizes $E[(G - \Gamma)^2]$.

We have

$$E[(G - \Gamma)^2] = \sigma^2 + B^2(N)$$

We therefore consider the problem of minimizing $\sigma^2 + B^2(N)$ under the restriction $\sum_{i=1}^{i=48} N_i = N$.

We have

$$\frac{\partial \sigma^2}{\partial N_i} = \frac{\partial \sigma^2}{\partial \rho_i}\frac{\partial \rho_i}{\partial N_i} = w_i^2 (1 - 2\rho_i) \frac{\partial \rho_i}{\partial N_i}$$

$$\frac{\partial B^2(N)}{\partial N_i} = 2B(N)\frac{\partial B(N)}{\partial N_i} = -2w_i B(N) \frac{\partial \rho_i}{\partial N_i}$$

$$\frac{\partial \rho_i}{\partial N_i} = -\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\gamma_i^2 N_i} \frac{\gamma_i}{2\sqrt{N_i}} \quad \text{if} \quad \pi_i < \tfrac{1}{2}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\gamma_i^2 N_i} \frac{\gamma_i}{2\sqrt{N_i}} \quad \text{if} \quad \pi_i > \tfrac{1}{2}.$$

Hence applying the method of Lagrange operators, we obtain

$$\text{(11)} \qquad \frac{\partial[\sigma^2 + B^2(N)]}{\partial N_i} = \frac{\partial \rho_i}{\partial N_i} w_i[w_i(1 - 2\rho_i) - 2B(N)] = \lambda, \quad i = 1 \cdots 48.$$

$$\sum_{i=1}^{i=48} N_i = N.$$

The parameters γ_i and π_i in equation (11) can be estimated from a previous poll.[3] It is not certain that (11) has always solutions. However if the quantity $\sigma^2 + B^2(N)$ has a minimum for a set of values $N_1, \cdots, N_{48}$ with $N_i \neq 0$ ($i = 1, \cdots, 48$) then (11) must have a solution

One might be induced to try to estimate $\Sigma\, \rho_i w_i$ directly by using $r_i = \frac{1}{\sqrt{2\pi}} \int_{(\frac{1}{2}-p_i)/s_i}^{\infty} e^{-x^2/2}\, dx$ as an estimate of ρ_i. It is easy to see that r_i is a consistent estimate of ϵ_i. It will be shown however that this estimate is more biased than the estimate (1).

Since σ_i differs only very little from its sample estimate s_i we may replace this sample estimate by σ_i. We then have

$$\begin{aligned} E(r_i) &= E\left(\frac{1}{\sqrt{2\pi}\,\sigma_i} \int_{\frac{1}{2}}^{\infty} e^{-(x-p_i)^2/2\sigma_i^2}\, dx\right) \\ &= \frac{1}{2\pi\sigma_i^2} \int_{-\infty}^{+\infty} \left(\int_{\frac{1}{2}}^{\infty} e^{-(x-p_i)^2/2\sigma_i^2}\, dx\right) e^{-(p_i-\pi_i)^2/2\sigma_i^2}\, dp_i \\ &= \frac{1}{2\pi\sigma_i^2} \int_{-\infty}^{+\infty} \int_{\frac{1}{2}}^{\infty} e^{-[(x-p_i)^2+(p_i-\pi_i)^2]/2\sigma_i^2}\, dx\, dp_i . \end{aligned}$$

Now

$$(x - p_i)^2 + (p_i - \pi_i)^2 = \frac{(x - \pi_i)^2}{2} + 2\left(p_i - \frac{\pi_i + x}{2}\right)^2.$$

Hence

$$E(r_i) = \frac{1}{2\pi\sigma_i^2} \int_{\frac{1}{2}}^{\infty} e^{-(x-\pi_i)^2/4\sigma_i^2} \left(\int_{-\infty}^{+\infty} e^{-(p_i-\frac{1}{2}(\pi_i+x))^2/\sigma_i^2}\, dp_i\right) dx.$$

The second integral is equal to $\sqrt{\pi\sigma_i^2}$. Hence

$$E(r_i) = \frac{1}{2\sqrt{\pi\sigma_i^2}} \int_{\frac{1}{2}}^{\infty} e^{-(x-\pi_i)^2/4\sigma_i^2}\, dx = \frac{1}{\sqrt{2\pi}} \int_{(\frac{1}{2}-\pi_i)/\sigma_i\sqrt{2}}^{\infty} e^{-x^2/2}\, dx .$$

[3] If π_i for any i were very close to $\frac{1}{2}$ then it would be of little use to poll the ith state. Hence, in this case formula (11) gives a small value for N_i. However, the π_i are never accurately known. The following procedure might be recommended for determining the best distribution of interviews: If for one particular i the sample value of π_i as estimated from a previous poll is too close to $\frac{1}{2}$ determine, using the N_i of the previous poll, that value $\bar{\pi}_i$ of π_i for which the probability is $\frac{8}{10}$ that p_i is larger than $\frac{1}{2}$ and substitute in (11) $\bar{\pi}_i$ for π_i. In all other cases substitute the sample value.

If several polls are taken it is advisable to use all of them but the last one to estimate as closely as possible the values of the π_i. The sample of the last poll before the election should be distributed according to (11).

From (12) we see that $E(r_i) < \rho_i$ if $\pi_i > \frac{1}{2}$ and $E(r_i) > \rho_i$ if $\pi_i < \frac{1}{2}$. Thus in every case this estimate is more biased than the estimate (1).

On the other hand, we shall now show that $E[(\epsilon_i - r_i)^2]$ is always smaller than $E[(\epsilon_i - e_i)^2]$. Since $\epsilon_i = 1$ if $\pi_i > \frac{1}{2}$ and $\epsilon_i = 0$ if $\pi_i < \frac{1}{2}$ it is easy to verify that $E[(\epsilon_i - r_i)^2]$ has the same value for $\pi_i = a$ as for $\pi_i = 1 - a$ and the same is true for $E[(\epsilon_i - e_i)^2]$. We may, therefore, without loss of generality assume that $\pi_i < \frac{1}{2}$.

Thus we have to show that

$$(13) \qquad E(r_i^2) \le E(e_i^2) = \rho_i = \int_{(\frac{1}{2}-\pi_i)/\sigma_i}^{\infty} e^{-\frac{1}{2}x^2}\,dx \qquad \text{if } \pi_i < \tfrac{1}{2}.$$

We have

$$E(r_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i}\int_{-\infty}^{+\infty} e^{-(p_i-\pi_i)^2/2\sigma_i^2}\left(\int_{(\frac{1}{2}-p_i)/\sigma_i}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\,dx\right)^2 dp_i$$
$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\int_{\frac{1}{2}}^{\infty}\int_{\frac{1}{2}}^{\infty}\frac{1}{2\pi\sigma_i^3}e^{-(1/2\sigma_i^2)Q(x,y,p_i)}\,dx\,dy\,dp_i\,.$$

Now

$$Q(x, y, p_i) = (x - p_i)^2 + (y - p_i)^2 + (p_i - \pi_i)^2$$
$$= 3\left(p_i - \frac{x + y + \pi_i}{3}\right)^2 + \frac{1}{6}(x + y - 2\pi_i)^2 + \frac{1}{2}(x - y)^2.$$

Putting

$$p_i' = \frac{\sqrt{3}\left(p_i - \dfrac{x + y + \pi_i}{3}\right)}{\sigma_i}, \quad x' = \frac{1}{\sqrt{6}}\frac{(x + y - 2\pi_i)}{\sigma_i},$$
$$y' = \frac{1}{\sqrt{2}}\frac{(x - y)}{\sigma_i}, \quad \frac{1 - 2\pi_i}{\sqrt{6}\sigma_i} = a,$$

we obtain

$$E(r_i^2) = \frac{1}{(\sqrt{2\pi})^3}\int_{-\infty}^{+\infty}\int_a^{\infty} e^{-\frac{1}{2}p'^2}e^{-\frac{1}{2}x'^2}\left(\int_{\sqrt{3}(a-x')}^{\sqrt{3}(x'-a)} e^{-\frac{1}{2}y'^2/2}\,dy'\right)dx'\,dp'$$
$$= \frac{1}{2\pi}\int_a^{\infty'} e^{-\frac{1}{2}x^2}\int_{\sqrt{3}(a-x)}^{\sqrt{3}(x-a)} e^{-y^2}\,dy\,dx.$$

Now for $\pi_i = \frac{1}{2}$ we have $a = 0$, and for $\pi_i < \frac{1}{2}$ we have $a > 0$. For $a = 0$ we obviously have $E(r_i^2 \le E(e_i^2)$. Further $\lim_{a\to\infty} E(r_i^2) = \lim_{a\to\infty} E(e_i^2) = 0$ hence (13) is proved if we can show that

$$F(a) = E(r_i^2) - E(e_i^2) = \frac{1}{2\pi}\int_a^{\infty} e^{-\frac{1}{2}x^2}\int_{\sqrt{3}(a-x)}^{\sqrt{3}(x-a)} e^{-\frac{1}{2}y^2}\,dy\,dx - \frac{1}{\sqrt{2\pi}}\int_{\sqrt{\frac{3}{2}}a}^{\infty} e^{-\frac{1}{2}x^2}\,dx$$

is a monotonically increasing function of a. Differentiating $F(a)$ with respect to a we obtain

$$\frac{dF(a)}{da} = -\frac{\sqrt{3}}{\pi}\int_a^\infty e^{-\frac{1}{4}(4x^2-6ax+3a^2)} + \frac{\sqrt{3}}{2\sqrt{\pi}}e^{-(3/4)a^2}$$

$$(14) \qquad = \frac{-\sqrt{3}}{\pi}e^{-(3/4)a^2}\int_a^\infty e^{-(x-(3/4)a)^2}\,dx + \frac{\sqrt{3}}{2\sqrt{\pi}}e^{-(3/4)a^2}$$

$$= \frac{-\sqrt{3}}{2\pi}e^{-(3/4)a^2}\int_{\frac{a}{2}}^\infty e^{-\frac{1}{2}x^2}\,dx + \frac{\sqrt{3}}{2\sqrt{\pi}}e^{-(3/4)a^2}.$$

Hence for $a \geq 0$ we have

$$\frac{dF}{da} \geq \frac{-\sqrt{3}}{2\sqrt{2\pi}}e^{-\frac{3}{4}a^2} + \frac{\sqrt{3}}{2\sqrt{\pi}}e^{-\frac{3}{4}a^2} \geq 0.$$

Hence we have proved

$$(15) \qquad E[(\epsilon_i - r_i)^2] = \frac{1}{2\pi}\int_{|a|}^\infty e^{-\frac{1}{2}x^2}\int_{\sqrt{3}(|a|-x)}^{\sqrt{3}(x-|a|)} e^{-y^2}\,dy\,dx \leq E[(\epsilon_i - e_i)^2],$$

$$a = \frac{1-2\pi_i}{\sqrt{6}\,\sigma_i}.$$

Since

$$E[(\epsilon_i - e_i)^2] - E[(\epsilon_i - r_i)^2]$$

is largest when $\pi_i = \frac{1}{2}$ we also have

$$E[(\epsilon_i - r_i)^2] \geq |\epsilon_i - \rho_i| - \left[\frac{1}{2} - \frac{1}{2\pi}\int_0^\infty e^{-\frac{1}{2}x^2}\int_{-\sqrt{3}x}^{+\sqrt{3}x} e^{-\frac{1}{2}y^2}\,dy\,dx\right]$$

or

$$(16) \quad |\epsilon_i - \rho_i| \geq E[(\epsilon_i - r_i)^2] \geq \frac{1}{2\pi}\int_0^\infty e^{-\frac{1}{2}x^2}\int_{-\sqrt{3}x}^{+\sqrt{3}x} e^{-\frac{1}{2}y^2}\,dy\,dx - |\tfrac{1}{2} - \rho_i|.$$

Because of (15), r_i although more biased may in many cases be preferable to e_i as an estimate of ϵ_i.