Understanding the exit poll calculations

Herbert Engstrom, Ph.D.
Tabor Enterprises, 5974 Friar Way, San José CA 95129, engstrom@best.com
November 2004

There is a number floating around the Internet, 250 million to one. This is alleged to be the chance that the exit polls in Florida, Ohio, and Pennsylvania could all be wrong. For each state the exit polls predicted that Kerry would gain a substanitally larger number of votes than were actually reported officially, and that he would in fact win both Pennsylvania and Ohio (and therefore the presidency) and lose Florida by a very small margin. The official tally awarded Kerry both Florida and Ohio by fairly large margins. The 250 million to one number was calculated by Steven F. Freeman of the University of Pennsylvania and appeared in an article entitled "The Unexplained Exit Poll Discrepancy." What Freeman calculated was the probability of those disparities for each of the three states, and, making the reasonable assumption that these probabilities were independent of each other, he multiplied these to calculate the probability that all three results could happen. The result was one chance in 250 million. Freeman did not conclude that either that the exit polls were wrong or that the vote tallies were wrong—only that the disparity requires further investigation.

To better understand the assumptions that Freeman made I have repeated his calculation for one state, that of Ohio, and I show in this note in some detail how the calculation was done. I also show how margin of error is calculated. The calculation is somewhat technical, but for those interested I present the results here.

Remember that if we have N possible, equally likely outcomes of an event such as drawing a card from a deck, and M of those are favorable (drawing an ace, for example) then we *define* the probability of this favorable event to be

$$p \equiv \frac{M}{N}. \tag{1}$$

Thus, for example, since there are 4 aces in the deck, $M = 4$, and 52 cards, $N = 52$, the probability of drawing an ace is

$$p = \frac{4}{52} = \frac{1}{13}.$$

Although Eq. (1) is a definition, it turns out to work in practice.

Suppose you perform n independent trials of something such as flipping a coin n times or drawing a card from a deck n times. By "independent" we mean that the result of each trial does not depend on any of the previous trials. In the case of drawing a card, it means we replace the card each time and shuffle the deck. Suppose the probability of a "favorable" event one each trial (getting heads, 1/2, or drawing an ace, 1/13) is p. Then the probability of x favorable outcomes in those n trials is given by the binomial frequency function:

$$B(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x} \tag{2}$$

where

n = total number of trials (e.g., coins flipped)
$n! = n \cdot (n-1) \cdots 3 \cdot 2 \cdot 1$
x = number of favorable events (e.g., heads)
p = probability of a favorable event (e.g., $p = 0.5$ for heads)
$q = 1 - p$ = probability of the unfavorable event (e.g., tails)

If n is large (say, 20 or more) we can use Stirling's approximation:

$$n! \approx n^n e^{-n} \sqrt{2\pi n} \tag{3}$$

to show that

$$B(x) \approx p(x) = \frac{1}{\sqrt{2\pi npq}} e^{-(x-np)^2/2npq}. \tag{4}$$

Now define

$$\mu \equiv np, \tag{5}$$

and

$$\sigma \equiv \sqrt{npq}. \tag{6}$$

We find

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}. \tag{7}$$

Statisticians call probabilities that have this form a "normal distribution."

The peak of the distribution occurs at $x = \mu$, and σ is a measure of the width of the distribution. The value of $p(x)$ falls to about 61% of its peak value at $x = \mu \pm \sigma$.

Eq. (7) gives the probability only when x is an integer as in the number of heads in a coin toss. More generally we consider the case when x is a random variable from a continuous domain. Examples are most commonly the result of a an experimental measurement of some physical parameter. In such cases we define a probability density, $f_x(x)$, and to calculate a probability we must specify an interval dx, which should be small enough that $f_x(x)$ does not change appreciably in that interval. We write that probability as

$$f_x(x)\, dx = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx. \tag{8}$$

If x is such a continuous random variable, then the question, "What is the probability that $x = 120$ when $\mu = 100$ and $\sigma = 15$?" has no meaning. The question: "What is the probability that x lies between 119 and 121 when $\mu = 100$ and $\sigma = 15$?" has the answer

$$p(x)\, dx = \frac{1}{15\sqrt{2\pi}} e^{-(120-100)^2/2\cdot 15^2} \cdot (121 - 119) = 0.0219.$$

The values that Freeman quotes for the Ohio result are as follows: n = sample size = 1963. This is the number of voters exiting the polls that were asked how they voted. The percent of those indicating a vote for Kerry was 52.1%, but the actual tallied vote was only 48.5%.

I have a slight quibble with Freeman. He assumed the exit poll was accurate and then calculated the probability that the actual vote could be as low as 48.5%. I believe the correct question is this: Given that the tabulated percentage that Kerry received was only 48.5%, what is the probability that the exit poll would give Kerry a vote at least as great as 52.1%? The calculation assumes that we have a valid statistical sample (a rather big "if") and that all assumptions underlying the normal distribution are also valid. In fact the way Freeman seems to have done the calculation and the way I will do it are essentially identical, and even the numerical results turn out to be the same. So here goes.

We find from (5) that the mean number of Kerry voters we expect in our sample, if the voter tally for the state is correct, is

$$\mu = np = 1963 \times 0.485 = 952, \tag{9}$$

where I have rounded to an integer, since people normally are counted as such. The standard deviation, σ, need not be an integer and is given by (6)

$$\sigma = \sqrt{np(1-p)} = \sqrt{1963 \times 0.485 \times 0.515} = 22.143. \tag{10}$$

Now if the exit poll predicted a Kerry vote of 52.1%, the number of Kerry voters in that poll must have been from (1)

$$m = np_m = 1963 \times 0.521 = 1022.7 \tag{11}$$

To be conservative, I'll round that number down so that

$$m = 1022. \tag{12}$$

The probability that Kerry gets this number or larger is from (8)

$$Q(m) = \frac{1}{\sigma\sqrt{2\pi}} \int_m^\infty e^{-(x-\mu)^2/2\sigma^2} dx. \tag{13}$$

Fig. 1 shows what the function looks like.

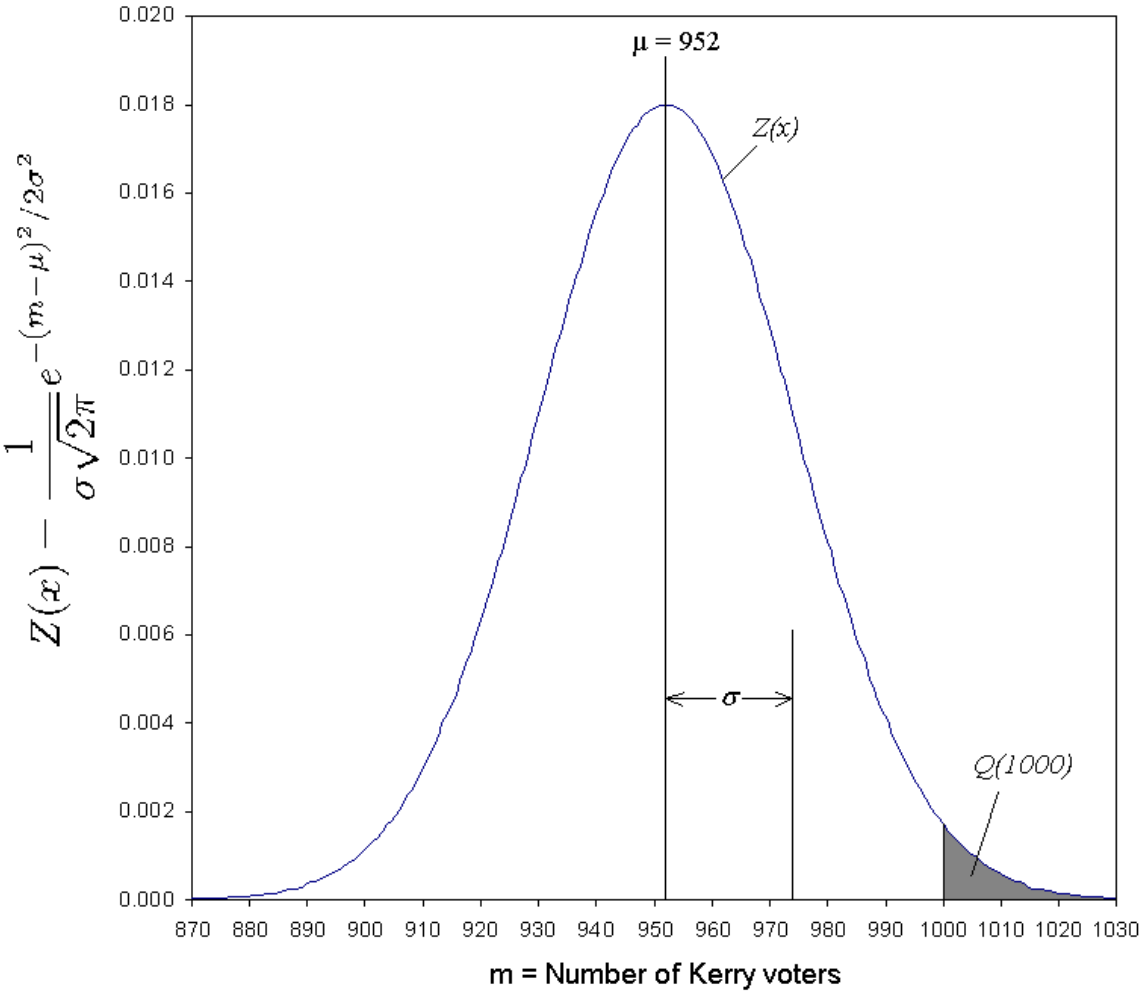

Fig. 1. $Z(x)$ is the probability distribution. $Q(m)$ is the area under the curve, $Z(x)$, for $x = m$ to $x = \infty$. The shaded area shows $Q(m)$ for $m = 1000$.

There is no exact analytical expression for $Q(m)$, but there are a number of highly accurate approximations. We appeal to the bible such approximations, which is the *Handbook of Mathematical Functions* by Milton Abramowitz and Irene A. Stegun published by the National Bureau of Standards. A slight modification of their equation 26.2.17 leads to

$$Q(u) = \frac{1}{\sqrt{2\pi}} \int_u^\infty e^{-u'^2/2} du' = Z(u)(b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5) + \epsilon(u), \tag{14}$$

for $0 \leq u < \infty$ where

$$t = \frac{1}{1 + cu}, \tag{15}$$

and

$$c = 0.2316419,$$

and

$$|\epsilon(u)| < 7.5 \times 10^{-8}$$

is the maximum error. Also in Eq. (14)

$$Z(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \tag{16}$$

and the coefficients are

$$\begin{aligned} b_1 &= 0.319381530, \\ b_2 &= -0.356563782, \\ b_3 &= 1.781477937, \\ b_4 &= -1.821255978, \end{aligned}$$

and

$$b_5 = 1.330274429.$$

Although (14) applies only for $0 \leq u < \infty$, because of the symmetry of (16), (14) can be modified easily for $-\infty < u < 0$.

To apply (14) to our case we need to make (13) look like (14). To do that, we define

$$u(m) \equiv (m - \mu)/\sigma. \tag{17}$$

Using our values of μ, σ, and m from (9), (10), and (12), we can apply (14) to find

$$Q(1022) = 0.00079, \tag{18}$$

identical to the result within round off error (0.0008) that Freeman obtains.

There is one more topic that is useful to understand—that of the "margin of error." When exit poll numbers are cited by the news organizations on election day, you will hear something like "In Ohio the exit poll predicts that Kerry will win with 52.1% of the vote. This number has a margin of error of 2.2%." Never do they tell you what that 2.2% means and how they arrive at it. Here's what and how.

There is a standard (or actually several standards) that statisticians use to estimate how accurately a number is known. What the margin of error tells you is that if the exit polls say Kerry will win with 52.1% with a margin of error of 2.2%, the actual tally will be between 52.1% − 2.2% = 49.9% and 52.1% + 2.2% = 54.3% with a 95% probability. That number, 95%, is called the "confidence interval." It is somewhat arbitrary but is commonly used; another commonly used confidence interval is 99%, but the margin of error reported in the news refers, I believe, to the 95% value.

To be more general, suppose the exit poll finds that μ people indicate they will vote for Kerry out of the sample of n voters polled. We need to find the number Δm voters such that the actual tally, μ' is such that

$$\mu - \Delta m \leq \ \mu' \leq \mu + \Delta m, \tag{19}$$

with a probability of $P = 95\%$. This is equivalent to asking, what is the value of

$$m = \mu + \Delta m \tag{20}$$

such that the probability that

$$m > \mu + \Delta m \tag{21}$$

is

$$v = (1 - P)/2 \tag{21}$$

or 2.5%? The answer is found from (13):

$$v = \frac{1}{\sigma\sqrt{2\pi}} \int_m^\infty e^{-(x-\mu)^2/2\sigma^2} dx. \tag{22}$$

This time, however, we know v and we need to solve (22) for m.

Again we appeal to Abramowitz and Stegun where we find their Eq. 26.2.3 is, with a change in notation:

$$v = Q(u) = \frac{1}{\sqrt{2\pi}} \int_u^\infty e^{-u'^2/2} du'. \tag{23}$$

Again, this equation is identical to (22) providing we again define

$$u = (m - \mu)/\sigma. \tag{24}$$

Equation (23) has the approximate solution given by Abramowitz and Stegun Eq. 26.2.23 for the case $0 < v \leq 0.5$:

$$u_v = h(t) + \epsilon(v). \tag{25}$$

where

$$h(t) = t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}, \tag{26}$$

$$t = \sqrt{\ln\left(\frac{1}{v^2}\right)}, \tag{27}$$

and

$$|\epsilon(v)| < 4.5 \times 10^{-4}. \tag{28}$$

The coefficients are given by

$$\begin{array}{ll} c_0 = 2.515517 & d_1 = 1.432788 \\ c_1 = 0.802853 & d_2 = 0.189269 \\ c_2 = 0.010328 & d_3 = 0.001308 \end{array}$$

Substituting $v = 0.025$ into (27) and t into (26) yields

$$u_{0.025} = 1.960 = (m - \mu)/\sigma = \Delta m/\sigma, \tag{29}$$

from which we find

$$\Delta m = u_{0.025}\sigma. \tag{30}$$

For the Ohio numbers we have

$$\sigma = \sqrt{np(1-p)} = \sqrt{1963 \times 0.521 \times 0.479} = 22.133, \tag{31}$$

and so

$$\Delta m = 1.960 \times 22.143 = 43.381. \tag{32}$$

Expressed as a percent,

$$\text{Margin of error} = \frac{\Delta m}{n} = \frac{43.381}{1963} = 0.0221 = 2.2\%. \tag{33}$$