TR2409-1: Polling 101 – Switch Probabilities

George Rebane – 10 September 2024

We must always remember that more often than not the public mind is fickle, frantic, and foolish.

Abstract – This report develops the argument that polls reporting the reliability of their results with something called 'within the margin of error' communicate almost nothing about the reliability of its stated numerical outputs of the reported preferences. The primary information about the reliability of a poll's results answers how confident can we be about the order of the stated preferences. If given the choice between propositions, candidates, or products A and B, how reliable is, say, the poll's report that A% < B%? What we really want to know is the probability that in the sampled target population the actual preferences are A% > B% - that in reality the results are NOT switched so that the drawn conclusion from the poll is valid. We arge that this so-called 'switch probability' or its complement is what should be reported with polls preference results rather than the meaningless 'margin of error' measure that supports no realistic decision based on the poll. In the following we summarize the development of switch probabilities and examine the reliability of polling results under various combinations population preference fractions (percents) of the presented alternatives. A method is developed for the lay reader to calculate switch probability from the reported margin of error. For the interested reader a technical report (TR2409-1) is available that covers this development in more detail and examines an expanded set of polling results. An early introduction to the notion of poll results switching was presented as a preamble to the 2016 presidential election. 'Polling Phollies' the 4sep16 post on *Rebane's Ruminations* developed the notion that ranked poll results may actually be reversed by the overlap of the candidates' preference distributions.

1. Background - 'Tis again the polling season. We Americans are devoted to political polling results to tell us who or what is preferred by whom here and there. But very few of us know the nature of the polling beast, and what it is and is not telling us. Public opinion is a volatile phenomenon, it can change in an hour and most certainly does within several days given unfolding political events, statements, and media propaganda.

In this piece I want to cover just one very important aspect of interpreting poll results. Specifically, what are we to make of the frequent reports of two closely competing proposals or candidate preferences being within 'the margin of error' from each other. Such a report basically tells us that that there is a good chance that the results are actually in the opposite order from those in the report – i.e. in the target population they are switched. What we'd really like to see are robust preference percentages that reliably communicate who or what is currently the in favored sentiment of the population polled. 'Margin of error' labels don't communicate, and instead leave us confused and more likely to ignore the poll results.

But the realworld is what it is, and tightly run races almost always yield closely spaced results (percentages) when comparing the preferences for candidate/proposition A versus B. Specifically, when A% > B% and the difference A% - B% is small, what we'd really like to have reported is the probability or odds that A really is preferred to B in the polled population. Or its opposite, that in reality B is more likely preferred to A regardless of the numerical percentages reported. In the sequel, I'll explain how such a useful measure of reliability, called the switch probability P_{AB}, may be calculated for any given poll (and should be calculated by the polling outfit). This development obtained from doing a little squiggly pushing and writing some software to calculate and plot the results – while not exactly rocket science, it is still important stuff.

2. A Detour on Distributions – The first thing to remember about any reported polling percentages that represent preferences is that they are random variables (RVs), and as all such RVs, they are 'drawn' from a probability distribution that determines which values are more likely to be drawn than others. Understanding this process is fundamental to understanding what a poll is and how to interpret its results. If you already understand probability distributions, you can safely skip this section, else read on.

In our universe ALL things observed, measured, estimated, … are RV – in the realworld nothing is precisely repeatable. Every time we look again, we get a different value with some that may look the same only if our measurements are sufficiently imprecise. That means every look is another draw from the RV's underlying probability distribution. The distributions of interest to us come in two flavors – continuous functions (curvy line) and histograms (bar charts).

Most distributions in our universe resemble the ubiquitous and familiar bell curve - more or less. These have a single prominent hump about which they may be symmetric or skewed. The relative height of the distribution (function) indicates the relative likelihood of drawing its RVs with the underlying value – e.g. RVs from a bell curve cluster mostly in the region of its hump (or mode, the distribution's maximum value), therefore the average value (or mean) of RVs from such distributions are almost always located under the hump. The mean and mode represent two measures of a distribution's 'central tendency'; the third is its median value, the 50-50 point located so that the probabilities of drawing an RV from either side of the median are equal.

The other important characteristic of a distribution is its 'dispersion', or how bunched or spread out it is over the domain of its RVs. The prime measures of dispersion are the distribution's standard deviation and variance, which is the square of the standard deviation. To best understand dispersion we need to be familiar with how probability distributions let us calculate actual probabilities for the values or range of values its RVs may be drawn or express themselves. Probability distributions may be expressed in two ways. The one we will be dealing with is actually called a probability density function (pdf) as represented by, say, the bell curve. Of interest to us will be pdfs that have the familiar hump in the middle and then either terminate or tail off to either side of its RV domain.

Now, say, we have x, an RV restricted to take values over its domain bounded by x_{min} and x_{max}. It is certain – i.e. the probability is one - that x will take a value from somewhere within its domain. This reality is represented by the area under its pdf function equaling one or unity. In Figure 2-1 below we have a graphic to illustrate the how the pdf of a given poll is construed and generated. The realworld poll result that we obtain to determine the share of the target population that favors A can be viewed as one sample, from a distribution of such sampled shares over multiple identical polls, which depends on the target population's actual fraction f_A of members who prefer A, the number of respondents N_S that we include in the poll, and the number of polls N_{polls} that we conduct – i.e. simulate.

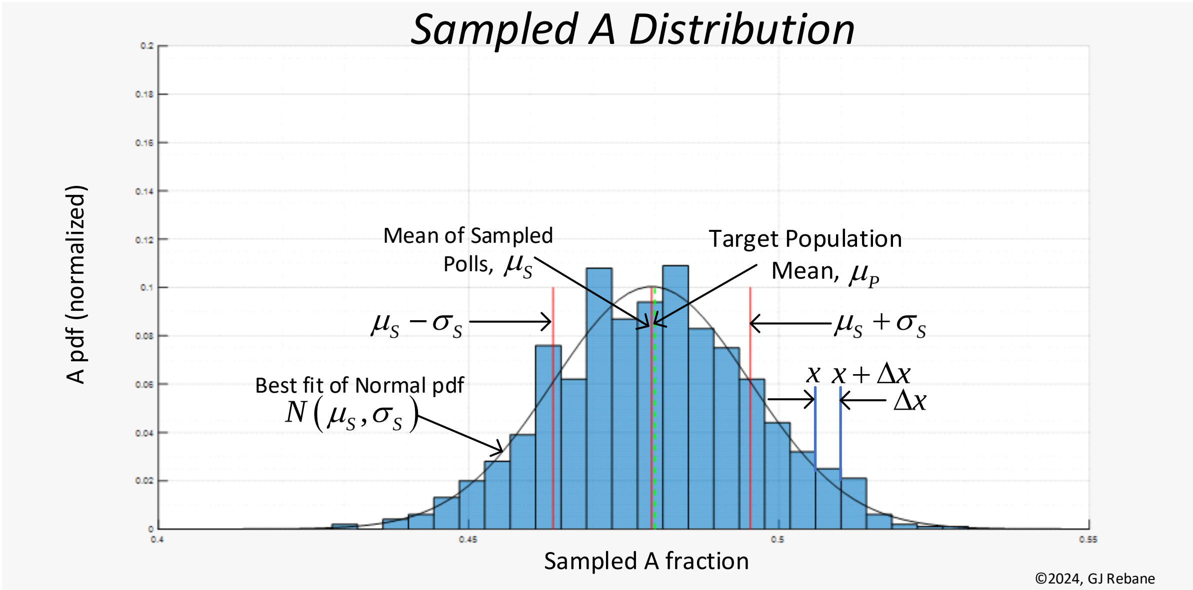

Figure 2-1

Making one such run of N_{polls} and plotting the resulting f_A fractions gives us the histogram in the figure. As shown, the height of each bar is proportional to the fraction values from the poll that fit within the histogram's bins of width Δx as x varies over its domain of values. The histogram's shape clearly resembles that of the single-humped normal pdf. If we normalize the bar heights to reflect the actual fractions of sampled responses falling into each Δx, then we can fit a smooth bell curve $N(\mu_S, \sigma_S)$ over the histogram that has the same mean μ_S and standard deviation σ_S of the all the respondents' data obtained from the simulated poll. Note that the target population mean μ_P (dashed green line) and the simulated mean μ_S are very close to each other and the input f_A to the simulation as expected.

When we repeat the experiment (simulate another poll) we will get a different histogram with somewhat different bar heights that still retains the shape, location, and width of the one shown. Finally, what is most important to note and understand is that each of the total N_{polls} simulated polls will yield a (slightly) different f_A value, as would happened were we to conduct multiple identical polls. Each of these fractions equals a poll's sample mean μ_S and, therefore, an estimate of the population's mean μ_P for the A cohort, and together they make up the new pdf (histogram) of the N_{polls} polls' distribution of means. Understanding the difference between a single poll's f_A distribution giving us a single μ_S from the poll's respondents (as shown in the figure) and the distribution of such poll means when all N_{polls} are simulated is critically important because it is the latter distribution (pdf) of poll means from which we will compute and generate the desired output switching probability of the actual poll that was conducted and reported.

3. Process Description - Now let's assume that in our polled population people are divided into three cohorts – those preferring A, those preferring B, and the undecided or independents preferring I (or possibly the aggregate of all other categories). The actual percentages of each cohort in the population are not known, and the objective of the poll is to determine these proportions. Based on the time available, polling costs, and other considerations, the pollsters decide on N_S, the number of people to include in their polling sample, and that the sample of respondents will be drawn through ‘raw sampling’ as opposed to, say, ‘demographically weighted sampling’ (q.v.) (In raw sampling every member of the target population has an equal chance of being included in the sample as a respondent.)

So we are interested in the preferences of a large target population numbering, say, $N_{pop} >> N_S$ members. Unknown to us there are the actual N_A, N_B, N_I members in each of the three cohorts comprising the surveyed population. In other words we have $N_A + N_B + N_I = N_{pop}$ which calculate to the actual fractional shares $f_A = N_A/N_{pop}$, $f_B = N_B/N_{pop}$, $f_I = N_I/N_{pop}$ that we wish to determine. To do this we sample the population generating a cohort of N_S respondents, the sample size. It is clear that each drawn respondent has the probability $P_A = f_A$, $P_B = f_B$, $P_I = f_I$, of belonging to each respective population cohort. Any raw sample of N_S that we select will divide into members N_{AS}, N_{BS}, N_{IS} such that it is highly likely that none of sample fractions $f_{AS} = N_{AS}/N_S$ etc where now $N_{AS} + N_{BS} + N_{IS} = N_S$ will equal the above actual population fractions f_A etc.

The best we can hope for is that our drawn sample will yield sample fractions that closely approximate the population fractions. Furthermore, we would like to determine how reliably can we conclude that $f_A > f_B$ when our poll outputs $f_{AS} > f_{BS}$ (or vice versa) with the available difference $f_{AS} - f_{BS}$. How this is done is covered in the sequel.

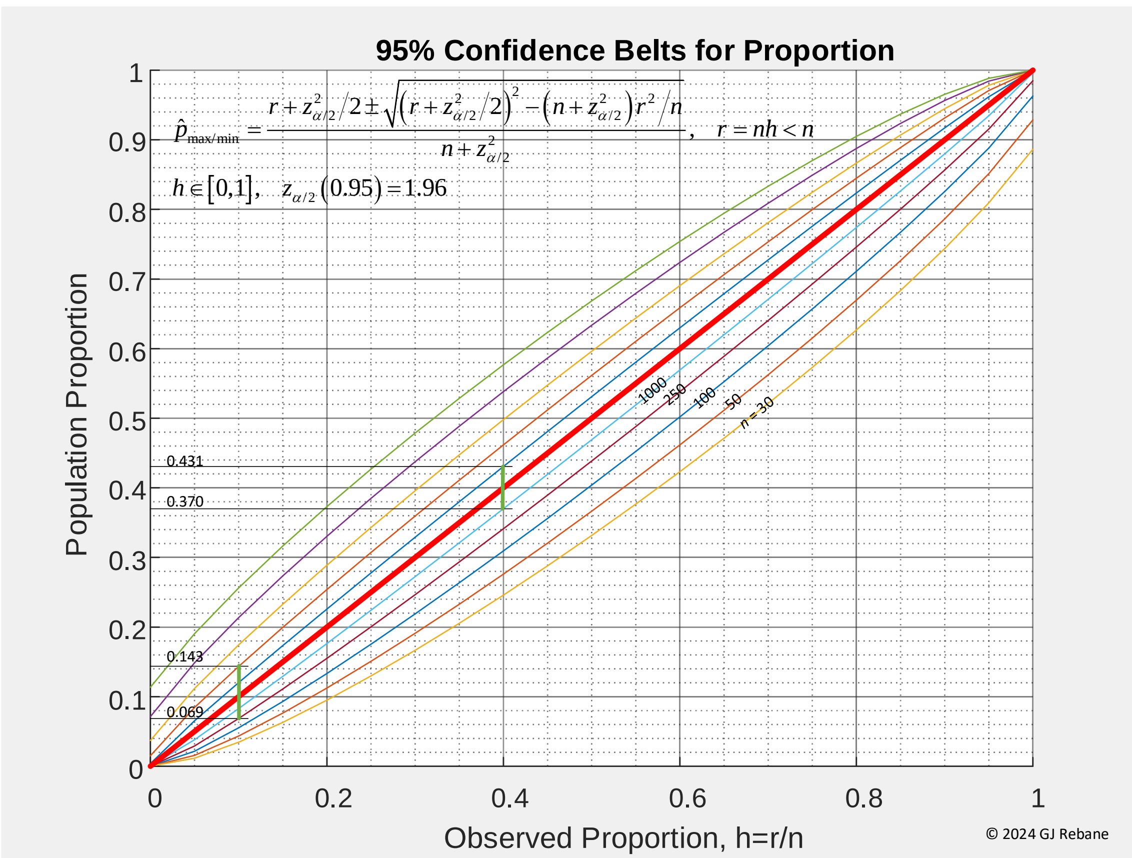

Figure 3-1

Before finishing this detour, we introduce the celebrated 'banana curves' in Figure 3-1 above that quantify the oft-quoted confidence intervals for the real population fractions at the 0.90, 0.95, 0.99 levels of probability (confidence). It is from these confidence intervals that the cited margins for error are derived.

The horizontal axis represents the observed poll fraction or proportion for, say, preference A. The vertical axis represents the values of that proportion in the sampled population. The curves come in pairs as a function of sample size. For example from the figure if our poll outputs f_{AS} = 0.4 from a sample of 1,000 respondents, then the real proportion of the population that prefers A lies between 0.370 and 0.431 with probability 0.95. This gives us the indicated 95% confidence level (green line) and confidence interval about the no error (red) line as the sample size is increased to the actual population size. 'Within the margin of error' is approximately one half the confidence interval or (0.431 – 0.370)/2 = 0.031. The confidence interval for f_{AS} = 0.1 is obtained the same way.

Another way of understanding the confidence interval is that were we to lay the histogram and its pdf bell curve in Figure 2-1 sideways on the Figure 3-1 confidence interval (green line) with the f_{AS} = 0.4 value centered on the diagonal red line, then 95% of the area under the bell curve would be bounded by the 1,000 respondent sample size banana curves as shown. It is in that interval [0.370 to 0.431] that the actual population fraction f_A would be found 95% of the time were this poll to be repeated. It is clear from the figure and as expected, that if the random sample of respondents were to be reduced, then the 95% confidence interval would increase, and our confidence in the accuracy of the poll's f_{AS} would decrease.

4. Technical Development – As described above, any poll we take will yield the triplet (f_{AS}, f_{BS}, f_{IS}). For convenience and notational efficiency let the vector $\underline{f}_S$ represent the triplet.

$$\underline{f}_S = \left(f_{AS},\, f_{BS},\, f_{IS} \right) \tag{4-1}$$

We know that if it were possible to immediately carry out another poll, or to poll two concurrent and equally drawn samples, we would not get the same $\underline{f}_S$ values. But we would expect them to be near each other. However, we can simulate the results of having conducted this poll a large N_{polls} number of times and examine the distribution of the resulting $\underline{f}_S$ cohort fractions. Plotting the histograms for this set of results $\left\{ \underline{f}_{S,i} \mid i = 1, N_S \right\}$ we obtain approximations of the familiar bell curve from which the cohorts' normal/gaussian density mean and variance values can be calculated. By the central limit theorem of probability[1] these pdfs will serve as the estimated proxies for all such polls conducted over the target population.

Figure 4-1 shows what the concurrent distributions of three preference cohorts looks like from one simulated poll of 1,000 respondents from a population for which f_A = 0.37, f_B = 0.43, f_I =

[1] "The Central Limit Theorem (CLT) in probability theory states that when you take a large number of random samples from a population, the distribution of the sample means will be approximately normal (bell-curve shaped), regardless of the original population's distribution shape, as long as the sample size is large enough; essentially, the average of many random variables tends towards a normal distribution even if the individual variables themselves are not normally distributed."

0.20. It's immediately clear that due to their proximity – i.e. the difference f_A - f_B is small - the distributions for the A and B cohort preferences overlap, giving rise to the AB overlap distribution which will now dominate our interest.

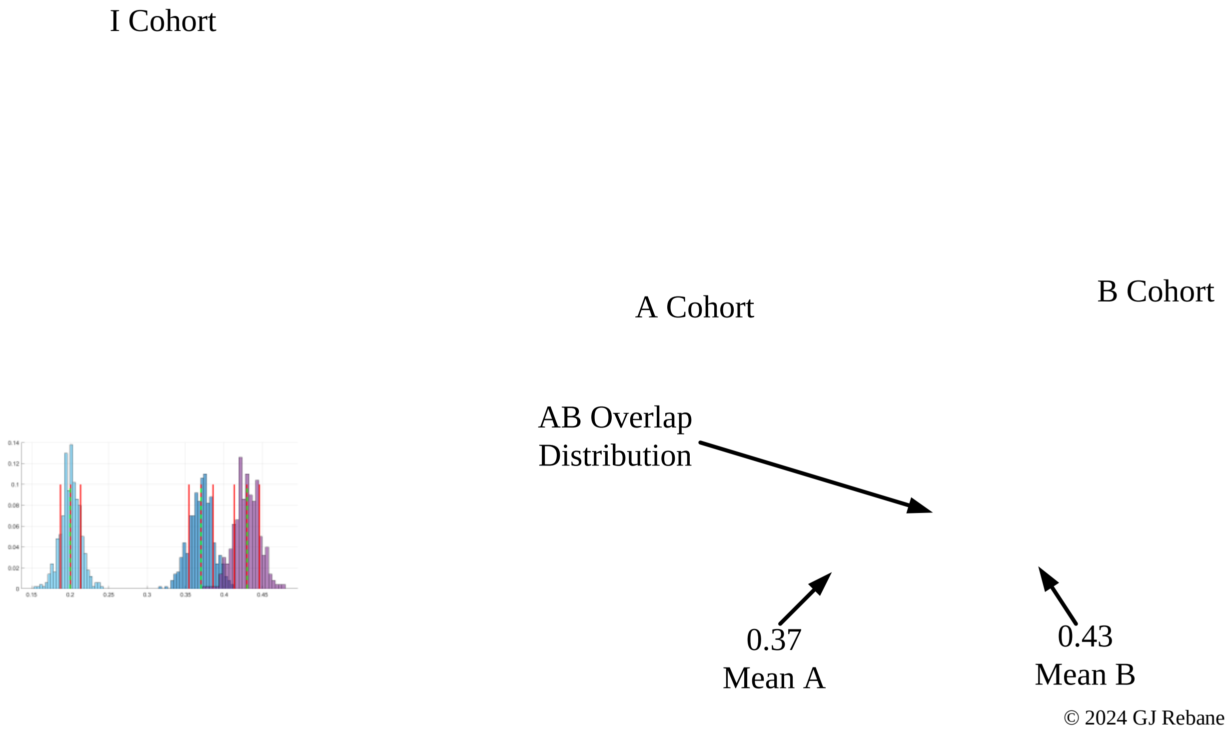

Figure 4-1

Since in the real world of polling we can only afford to take one poll, that poll should be viewed as a random sample drawn from the above-described cohort distributions with its set of cohort fraction means μ_{AS}, μ_{BS}, μ_{IS}. But since we can conduct N_{polls} through simulation, we can calculate a large set of these means $\{\mu_{ASi}, \mu_{BSi}, \mu_{ISi} \mid i = 1,N_{polls}\}$ and study their distributions using all the techniques that probability theory teaches, and derive our desired reliability measures.

(At this point it's important for the reader not to get lost in the weeds of all these distributions. From each simulated poll we get the preference fraction distributions of the respondents belonging to the prreference cohorts. From each of these fraction distributions we can calculate their distribution means that approximate the actual population means for the cohorts. From this set of means we obtain a new distribution of means from which we calculate the best estimate of the actual population cohort fractions for which the means serve as proxies. But more than that, with the cohort means distributions we can also generate other important distributions for functions of these means like that for the important difference μ_{Asi} - μ_{BSi}.)

We keep in mind that all the simulated fractions and means are RVs, random (and stochastic[2]) variables each with its own probability distribution (or pdf).

[2] Formally a stochastic variable is a function of one or more elementary random variables. If the RV is x = face value of a rolled die, then $4x^2$ is a stochastic variable with its own pdf. If rolling a three on a die and drawing a face

With these preliminaries in hand, we come to the main event, calculating the switch probability that a reported poll result f_A and f_B such that $f_A < f_B$ is really $f_A > f_B$ in the target population. We do this by forming a new RV with mean being the difference

$$\mu_{ABSi} = \mu_{ASi} - \mu_{BSi}, \quad i = \left[1, N_{polls}\right] \tag{4-1}$$

For this simulated dataset we compute the new normal distribution $N(\mu_{ABS}, \sigma_{ABS})$ where by the Central Limit Theorem we have

$$\mu_{ABS} = \frac{1}{N_{polls}} \sum_{i=1}^{N_{polls}} \mu_{ABSi}, \quad \mu_{AS} = \frac{1}{N_{polls}} \sum_{i=1}^{N_{polls}} \mu_{ASi}, \quad \mu_{BS} = \frac{1}{N_{polls}} \sum_{i=1}^{N_{polls}} \mu_{BSi}, \tag{4-2}$$

And (Papoulis p222, 4th ed p202)

$$\begin{aligned}
\sigma_{ABS}^2 &= \frac{1}{N_{polls}-1} \sum_{i=1}^{n} \left(\mu_{ABSi} - \mu_{ABS}\right)_i^2 \\
&= \left[\frac{\delta\mu_{ABS}}{\delta\mu_{AS}}, \frac{\delta\mu_{ABS}}{\delta\mu_{BS}}\right] \mathrm{cov}(\mathrm{AB}) \left[\frac{\delta\mu_{ABS}}{\delta\mu_{AS}}, \frac{\delta\mu_{ABS}}{\delta\mu_{BS}}\right]^{\mathrm{T}} = \begin{bmatrix}1 & -1\end{bmatrix} \begin{bmatrix} \sigma_{AS}^2 & \rho_{ABS}\sigma_{AS}\sigma_{BS} \\ \rho_{ABS}\sigma_{AS}\sigma_{BS} & \sigma_{BS}^2 \end{bmatrix} \begin{bmatrix}1 \\ -1\end{bmatrix} \\
&= \sigma_{AS}^2 - 2\rho_{ABS}\sigma_{AS}\sigma_{BS} + \sigma_{BS}^2
\end{aligned} \tag{4-3}$$

Here the second form of variance allows us to study the impact of possible correlations ρ_{ABS} between preferences A and B (especially since in a two preference poll their observed fractions tend to be correlated complements).

A poll's result is reported in the form of μ_A being less or greater than μ_B; or from (4-1) having μ_{AB} being greater or less than zero – i.e. positive or negative. That means that the switch probability P_{AB} is the cumulative probability on one or the other side of $\mu_{AB} = 0$. Without loss of generality we assume that the poll reports $f_A < f_B$, giving $\mu_A < \mu_B$ and a negative μ_{AB} value. Then from their cumulative distribution of $N(\mu_{AB}, \sigma_{AB})$, where now we use the proxies $\mu_{AB} \approx \mu_{ABS}$ and $\sigma_{AB} \approx \sigma_{ABS}$ which lets us write

$$\begin{aligned}
P_{AB} &= 1 - \int_{-\infty}^{0} N\left(\mu_{AB}, \sigma_{AB}\right) d\mu, \quad \text{for } \mu_{AB} < 0 \\
&= \int_{-\infty}^{0} N\left(\mu_{AB}, \sigma_{AB}\right) d\mu, \quad \text{for } \mu_{AB} \geq 0
\end{aligned} \tag{4-4}$$

These switch probabilities are illustrated below in Figure 4-2 for the two conditions of μ_{AB} on their normal distributions $N(\mu_{AB}, \sigma_{AB})$. The switch probabilities P_{AB} are the areas under the bell curve that is above the red line on the *x*-axis.

---

card results in a $5 payoff, then the money won after ten rolls is a stochastic variable with its own distribution. In this report we will continue using RV for both random and stochastic variables.

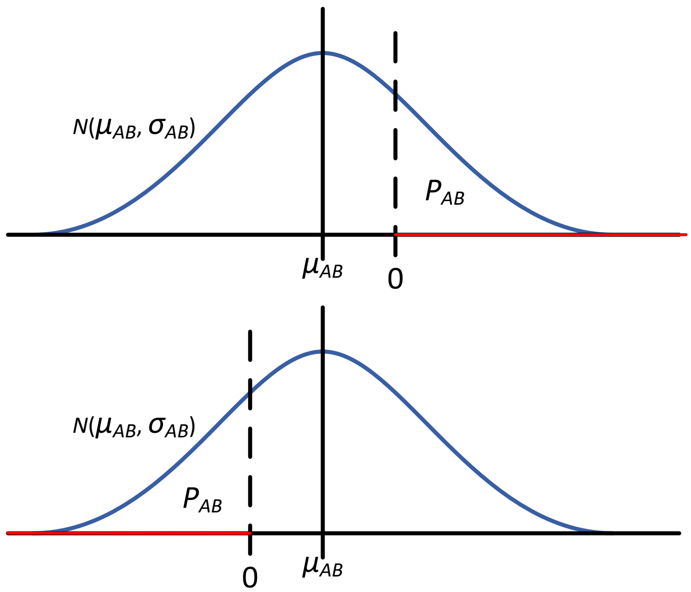

Figure 4-2

The reader can readily compute the switch probability for a poll for which the Margin of Error (MoE) is quoted as a percentage (converted to its decimal equivalent). As shown above, for the 0.95 confidence interval the margin of error equals about $2\sigma_{AB}$, therefore σ_{AB} equals MoE/2. The spreadsheet MS Excel™ is ideal for calculating P_{AB} from (4-4) by typing in the poll values for $|f_A-f_B|$ and MoE/2 in their decimal formats, and using these values in a cell with the formula

$$P_{AB} = \text{NORM.DIST}(0, |f_A-f_B|, \text{MoE}/2, \text{TRUE}) \tag{4-5}$$

which yields the desired P_{AB}. From Figure 4-2 we see that P_{AB} is actually symmetric with respect to the relative sizes of f_A and f_B, therefore allowing us to only one formula from (4-4), and also that the poll's cited MoE already takes into account the sample size of its respondents. Finally, in Figure 4-3 we present a table of switch probabilities for a typical range of $|f_A-f_B|$ and MoE values. For example, for $|f_A-f_B|$ = 2% and MoE = 4% we read P_{AB} = 0.1587 or a chance of about one out of six that the preference fractions in the target population are actually in the opposite order.

Switch Probabilities at 0.95 confidence level					
|fA-fB|\MoE	2%	3%	4%	5%	6%
0.5%	0.3085	0.3694	0.4013	0.4207	0.4338
1.0%	0.1587	0.2525	0.3085	0.3446	0.3694
1.5%	0.0668	0.1587	0.2266	0.2743	0.3085
2.0%	0.0228	0.0912	0.1587	0.2119	0.2525
2.5%	0.0062	0.0478	0.1056	0.1587	0.2023
3.0%	0.0013	0.0228	0.0668	0.1151	0.1587
3.5%	0.0002	0.0098	0.0401	0.0808	0.1217
4.0%	0.0000	0.0038	0.0228	0.0548	0.0912
4.5%	0.0000	0.0013	0.0122	0.0359	0.0668
5.0%	0.0000	0.0004	0.0062	0.0228	0.0478

Figure 4-3