ELECTRONOTES

WEBNOTE 17 2/13/2014

ENWN-17

Simple Illustrations of Statistical Significance

Recently, with regard to a local traffic study, I had occasion to wonder about how well the general public understood the notion of "statistical significance". As is often the case, when you imagine how you would go about explaining the concept to a less mathematically inclined person, you wonder if you too perhaps have an incomplete understanding of what is involved. Usually this turns out to be the case. Accordingly, perhaps careful artificial examples are one way to proceed. Specifically here we have as the issue measured speeds on a section of road before and after a particular change is made. There are many assumptions necessary to evaluate the data, and of course, a properly controlled experiment (not always easy) is required. And you need to collect data correctly. Then what do you do. Well, I think first you look at the data directly.

Not having any actual data, I decided to invent some test data based on what I knew I was likely going to have to assume anyway – a normal distribution. This I did in Matlab code as follows:

```
% make distributions of 10000 speeds centered at 50 and 55 std dev = 8
SD=8
SA=50
SB=55
A = SA+SD*randn(1,10000);
B = SB+SD*randn(1,10000);
```

This I supposed would set means of 50 and 55 (mph) with standard deviations of 8 mph. Since I seldom use ***randn***, I figured I would go right ahead and plot these, along with some parameters extracted from the generated data. These I show in Fig. 1.

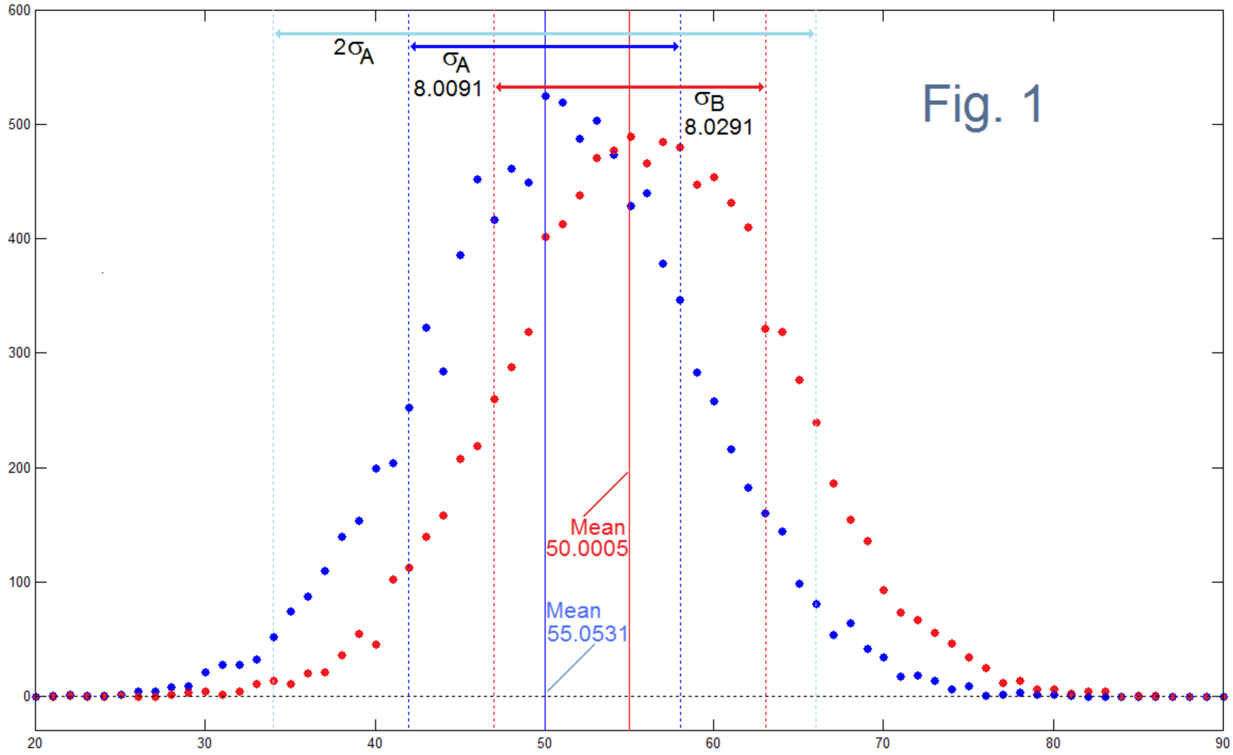


Fig. 1

I found Fig. 1 to be a surprise in two ways. First, these histograms seem to be bell curves, but I had expected that with 10,000 points I would see a much better-looking pair of curves. Secondly, the poor appearance notwithstanding, the mean (μ) and standard deviation calculations (symbol σ), from Matlab functions, are astoundingly good. We know what the "correct values" should be from the Matlab code that generated the data. Note how well these are "recovered" from the actual data points. One additional point we can check is how many data points are within 2σ which most of us perhaps remember should be about 95%. Here the blue curve was tested; it was 95.31%. [A bit more on this later, but here we also checked for being within 1.96σ (perhaps the "right" number) and this was 94.88%.] So far – so good.

CALCULATING FROM THE DATA

Here we started out knowing the mean and the standard deviation of the entire population (essentially infinite). because we typed the numbers into the Matlab code and then "sampled" 10,000 numbers of this background population. Then, ignoring the fact that we typed in these values, we calculated the mean and standard deviation from these 10,000 data points. Well. Matlab did this too – working with the sequences, and the ***mean*** and ***std*** functions. It helps to recognize that we could have calculated these by hand, given the data. Of course, we recall that the mean is just what is more usually called the average so:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (1)$$

And once we have found the mean from the data points x_i, we can solve for the standard deviation, or the RMS error (Root Mean Square). That is, we (1) compute the errors by subtracting the mean from the data points, then we (2) square the errors, (3) take the mean of the squared errors, and finally (4), the square root:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2} \qquad (2)$$

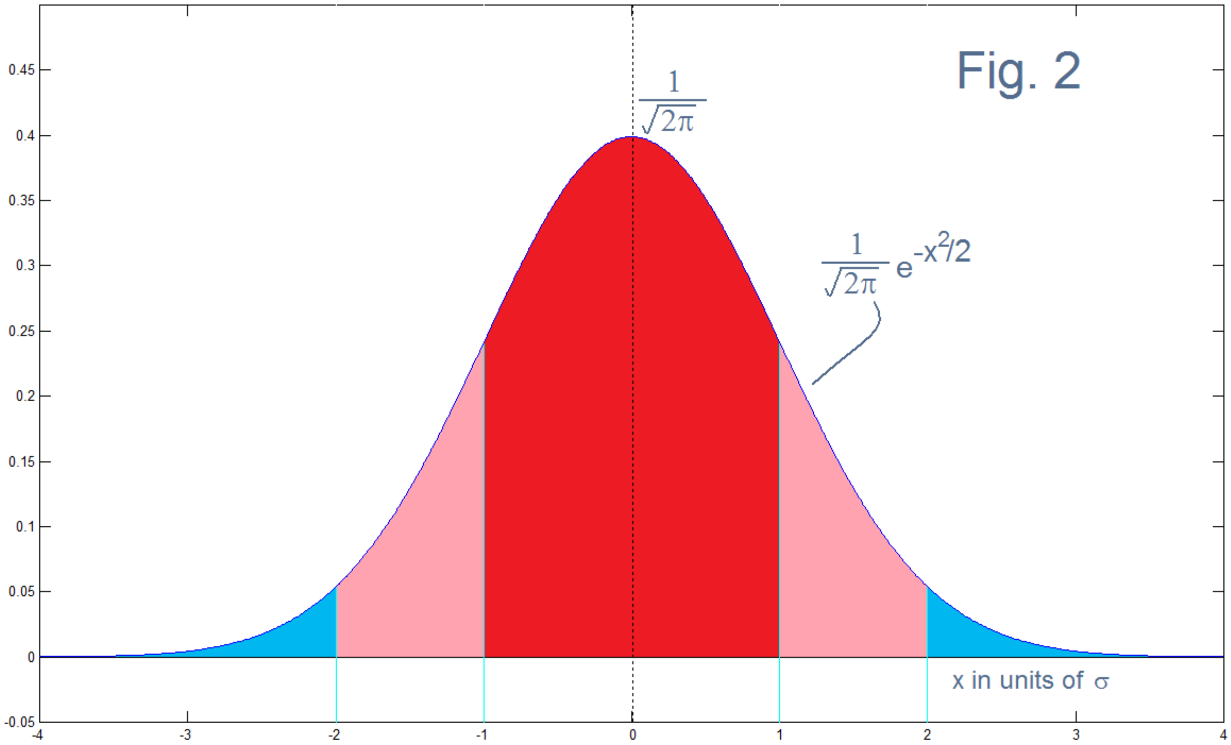
This would be tedious if not programmed. Many math languages will have standard deviation as a function. One subtle point is that when computing σ, sometimes we divide by N-1 instead of N. In many cases N is large enough that this makes little difference. Specifically, the case where we divide by N-1 is where we have N data points that comprise one sample instead of N being the whole population. This I do not really understand well. The "whole population" concept is easier. In the case of N students, their grades would represent the entire population. In the case of the traffic, it would seem that we have samples of a large population (at least including the cars we did not measure – the day before we put up the radar for example). But I am unsure how big this is. (All cars since the first Model-T?). For my simulations, I have N=1000 samples so will assume that is large enough that N or N-1 does not matter.

THE EQUATIONS

Equations (1) and (2) tell us how to calculate, given the data. Nothing about these equations requires that the data have a particular distribution such as "Normal" (also called Gaussian or Bell-Shaped). In the case where the data is, or is being assumed to be normal, the mean μ and the standard deviation σ will take on particular meaning. Fig. 2 shows the typical normal curve where we have set the mean to zero and the standard deviation to 1. The formula for this curve is:

$$x(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (3a)$$

which in this case simplifies to:



$$x(t) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad (3b)$$

This is a famous curve that has much magic and is probably quite capable of deception! Note that it is an idealized version of the two curves of Fig. 1 which are offset by their means, widened by their standard deviations (about 8) and which are represented by a “histogram” of points, and appear a bit irregular.

In Fig. 2, we show a graph width of ±4σ, which includes essentially the whole distribution. Possibly we recall that the total area under the curve has 68% within ±σ (red) and 95% within ±2σ (red + pink). Probably they did not tell us that these were approximate. (But we suspected that, and when talking statistics, most things are approximate!). But the area under the curve problem sure sound like a calculus problem, and we should be able to integrate from 0 to some value x’ for 95%/2.

$$\frac{1}{\sqrt{2\pi}} \int_0^{x\prime} e^{-x^2/2}\, dx = 0.475 \qquad (3b)$$

This is a trivial integral. Nope! You just think it should be trivial because it looks like an exponential.

I leave it to the reader to try to do it, then look for it in your tables, then look on the web as to why it is difficult or impossible. I did these three things. Then I just wrote a few lines of Matlab code to do it numerically.

```
% integrate normal

ii=0;
x=0;
aa=1/sqrt(2*pi);
while ii<0.95/2
    n=aa*exp(-x^2/2);
    ii=ii+n*.000001;
    x=x+.000001;
end
x
ii
```

The result is x = 1.9600 for the area of 0.4750. So we see the basis for the approximation that 95% of the area is within 2σ while the correct answer is 1.96σ as noted above.

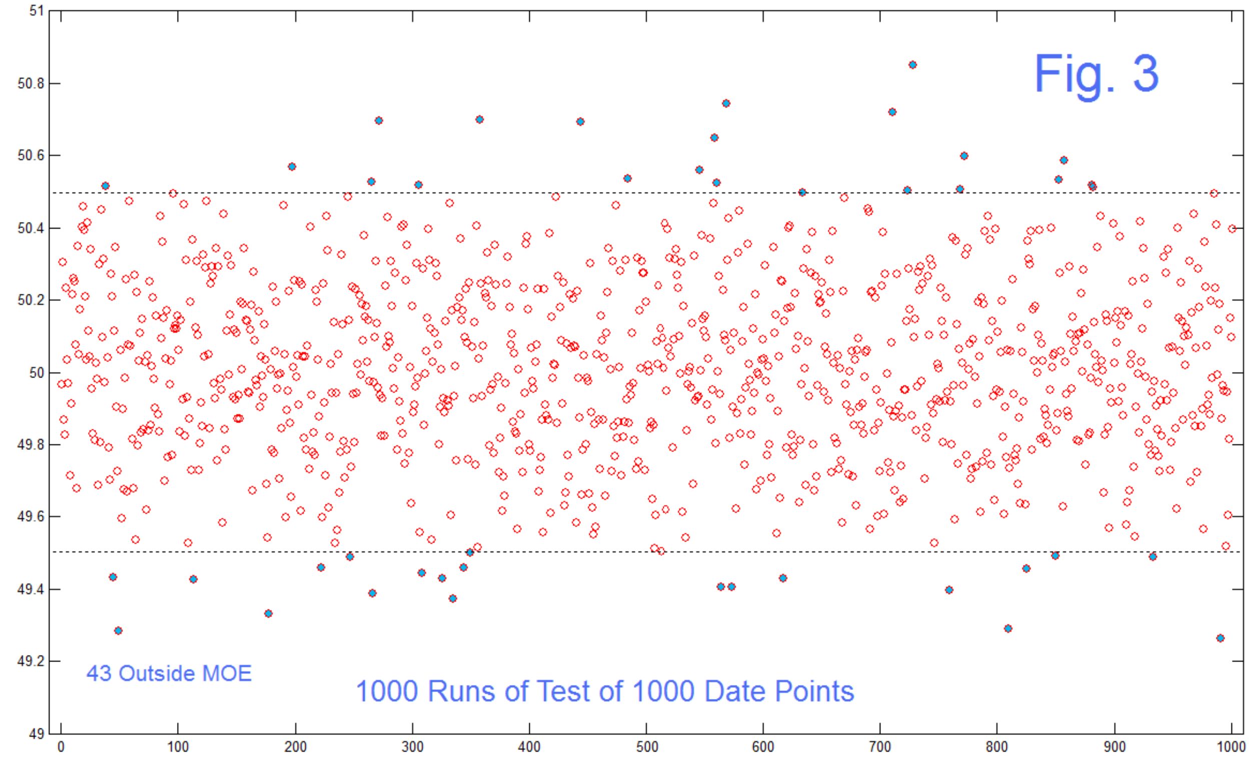
END REVIEW – WHAT ARE WE TRYING TO DO?

What we are trying to do is take the sample data set and calculate from it such things as the mean μ (which will be the main "output") and the standard deviation σ as side information. That is, we did our sampling and have that one sample. In general, probably this one sample is all we intend to take, for reasons of limited resources of time/money etc. (Keep in mind that a "sample" in statistics is not what it would be in signal processing; rather it is a whole set of data points, not the points themselves.)

Of the two main statistics from this one sample (μ and σ), μ is interesting, but σ is usually not part of the "public take-away". We want to find, from these, what the "confidence limits" are, or more famously the "margin of error". That is, how reliable is the number we calculated for the mean? So if we get from our data a mean of 52 mph based on the one "sample" of the N speed measurements we can afford, we want to know if we can be reasonably sure (typically 95% sure) that the "correct" mean of the original population is not (for example) 51 mph perhaps, or 55 mph perhaps.

Nothing prevents us from supposing we have unlimited resources and can take more samples. For our study the computer just makes up data and crunches it. Being so cheap, why not run 10,000 samples of length 1000 data points, and see what the means of these 10,000 trials look like. We could display these as a histogram, but here we will plot successive means for the first 1000 of the 10,000 runs. (We could plot all

10,000 but it would be too cluttered.) So Fig. 3 shows this plot of the first 1000 trials along with two doted black lines representing our Margin of Error (MOE), to be described. Note that 43 of the means (blue centers), counted by hand, are outside the MOE here.



The 43 means of 1000 speeds plotted would be 4.3%, respectably close to the 5% failures we were willing to accept (95% successes). The plot of Fig. 3 quite nicely <u>illustrates the idea</u> of finding means that happen to fall outside the MOE lines. For the full 10,000 means, we don't plot and count by hand, but rather the program counts these for us, and for the full 10,000 means the count here came out (embarrassingly!) to exactly 500. We feel obliged NOT to "censor" runs when doing a random study, so this exact 5% stays. Instead, we note that running additional trials of 10,000 means each, we get for the first ten: 500, 525, 500, 483, 508, 478, 490, 505, 500, and 499. At the moment, we feel we have found a good MOE for 95% inside. But where did the MOE, the dotted black lines of Fig. 3, come from? Here the MOE was calculated as:

$$MOE = 1.96\,(\;) \qquad (4)$$

where σ was set to 8 and N to 1000, so the MOE = 0.4958 mph (as in Fig. 3 where we see the confidence interval set between about 49.5042 mph and 50.4958 mph). Note that the "confidence interval" as ±MOE, is conventional. The term $\frac{\sigma}{\sqrt{N}}$ by itself is often called the "standard error".

The constant of 1.96 in equation (4) is approximately 2, and is sometimes written as such. If a confidence level other than 95% were chosen, the constant 1.96 in equation (4) changes. For example, for 90%, the constant becomes 1.6449 and for 99%, it becomes 2.5758 using the same basic code as that on the top of page 5, which we used to give us there the value 1.96.

We do not propose here to derive equation (4). Discussions of MOE are numerous, often emphasizing interpretation rather than origin, and get lost in additional terminology. Here we find careful study of Fig. 3 (as an imaginary exploitation of supposed immense resources) to be the next step. It is clear what Fig. 3 shows. There we know the true μ and the true σ. We find that for any one trial, the experimental μ is inside the MOE with probability 95% (or whatever has been chosen). That is, the difference between the experimental mean and the true mean is within the MOE.

Now, not knowing the true mean, we must estimate it as being the experimental value. (It’s all we usually have, and we worked hard just for that.) Indeed, we need this estimated μ, treated as the real μ, just to continue the processing to estimate σ and the MOE. The whole point is we need to estimate the true μ from the sample, as we don’t have the unlimited resources suggested by Fig. 3. So we turn around the reasoning that the experimental μ fall within the MOE 95% of the time to say that the true μ is within any experimental μ 95% of the time. That’s the interpretation from Fig. 3.

Thus an experimental mean of 50.2 mph (see Fig. 3 which has many trials of just about that value) is nicely within 0.4958 (the MOE, very close to 0.5 mph) of the true μ (50 mph but we don’t know this). We could also have had an experimental μ of 49.7 mph as the one we actually did get, again within the MOE of 50 mph. But of course, again from Fig. 3, at about trial 730 we found an experimental μ of about 50.83. Based on this, we would have had to estimate the true μ to be between 50.33 and 51.33, and we would be wrong. But we were supposed to be wrong 5% of the time. We don’t know, and can’t know without more effort, that we ARE in fact wrong. Of course, that is still a small error. The sample size of 1000 speeds pretty much rules out really bad experiments. We don’t see means in Fig. 3 of 60 mph.

This does not of course mean that no cars were going 60 mph. Indeed, we suspect that many were, as the normal curve with μ=50 and σ=8 tells us that we expect 2.5% of cars were going faster than 66 mph.

Note that traffic studies often cite an “85th Percentile” meaning a speed at which 85% of the vehicles are going slower, and 15% are going faster. Using again the code snippet at the top of page 5, we find that for a normal distribution, this would be at 1.44σ, which would be about 62 mph for our example. Most likely an 85th percentile would not be found from σ, but by a simple count through an ordered list of the measured speeds.

There are different presentations of MOE. I believe the one in equation (4) is the most useful as it provides the MOE with UNITS, specifically as whatever units we use to express σ. In our example case, it is mph. Sometimes it is discussed in terms of percentages, and typically we see a percentage of 3.1% for N=1000. (This is the usual poll result accompanying political standings, for example. SEE BELOW). This would give a MOE of 3.1%×2×8 = 0.496, the result we got above, where the 2×8 is two standard deviations, what we need for 95%. But percentages and "percentage points" can be confusing. Speaking of which…….

AND HOW ABOUT THAT ELECTION POLL?

We basically know that when candidate A has 55% and candidate B has 45% with a MOE of 3% that A is supposedly winning (everything else being perfect!). Yet what is missing in the usual media report is the actual meaning in terms of the probability of getting a particular result within certain confidence levels (within a confidence interval), which I think Fig. 3 provides. Equation (4) is not at all difficult to remember, especially if you think of 1.96σ as two sigma.

But something quite different jumps in when we consider an election poll. Usually you get to vote for, or express support for, only one candidate at the 100% level. In the traffic example, you might drive at a continuum of speeds, and a normal distribution seemed at least a realistic starting point. Here, if you are a candidate with 40% support, and we have N=1000, we envision a "distribution" that is zero everywhere except at 100% where it has value 400. No! We also have 600 at 0%. Two spikes (Fig. 4). Hardly a bell curve! So what could MOE mean? If we look for reference, we find a lot for this case, but about all they say is that the MOE is about 3.1% for 1000 people polled, and when they give a formula, it is $0.98/\sqrt{N}$ or just $1/\sqrt{N}$ which is about 0.031, the 3.1%.

So we hesitate to even think about a normal distribution. Can we still compute a mean and standard deviation. Well – yes. Equations (1) and (2) still work. However, so that we won't think too much about the missing bell curve, we had better call σ the "RMS Error" as we noted above already. So, the candidate has 400 voters at 100%. His mean support (Equation 1) gives (400×100%)/1000 = 40%. Of course. For the RMS error, we note that there are 600 points where the poll was 0% and 400 points where the poll was 100%. The mean being 40%, the errors are 40% and 60% respectively. So we take the square root of $[600 \times 40^2 + 400 \times 60^2]/1000$ which is about σ = 49%. Now $1.96\sigma/\sqrt{N}$ is 3.04%. So I guess this still works or at least agrees with what the reporter says!

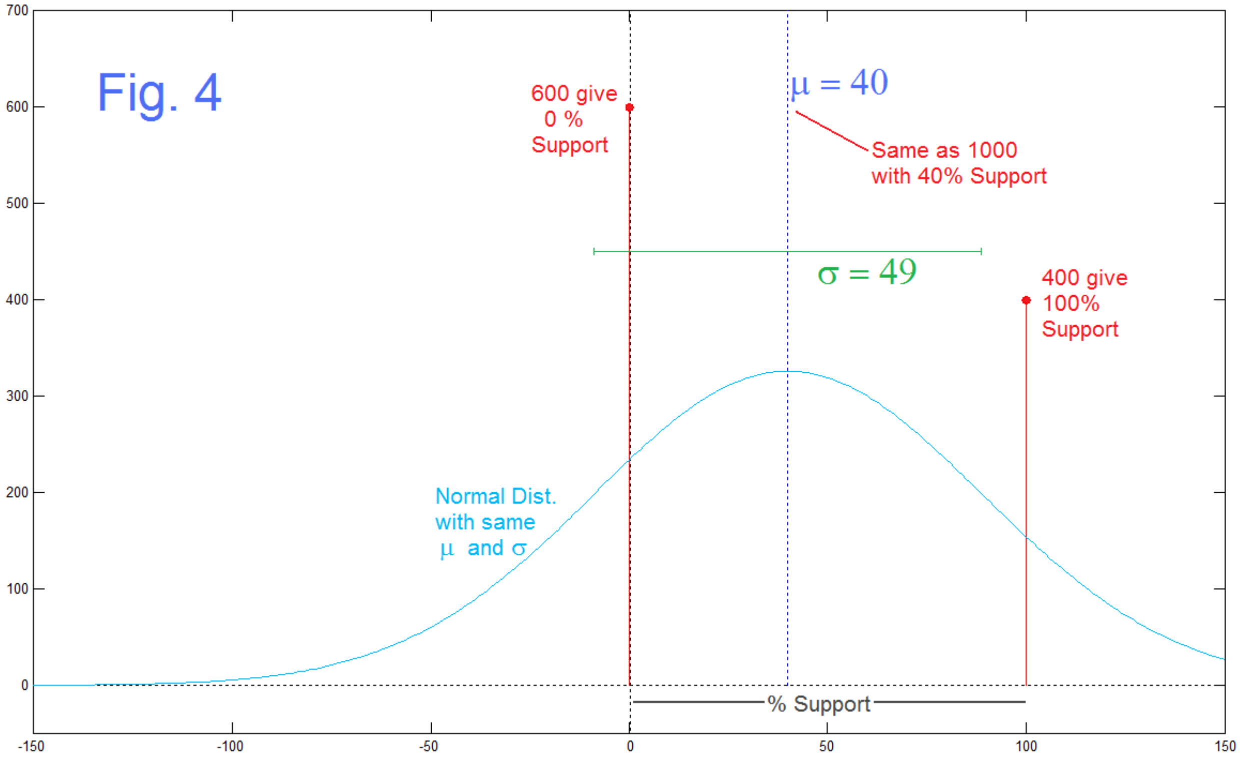


Fig. 4

700
600
500
400
300
200
100
0
-150
-100
-50
0
50
100
150
600 give
0 %
Support
μ = 40
Same as 1000
with 40% Support
σ = 49
400 give
100%
Support
Normal Dist.
with same
μ and σ
% Support

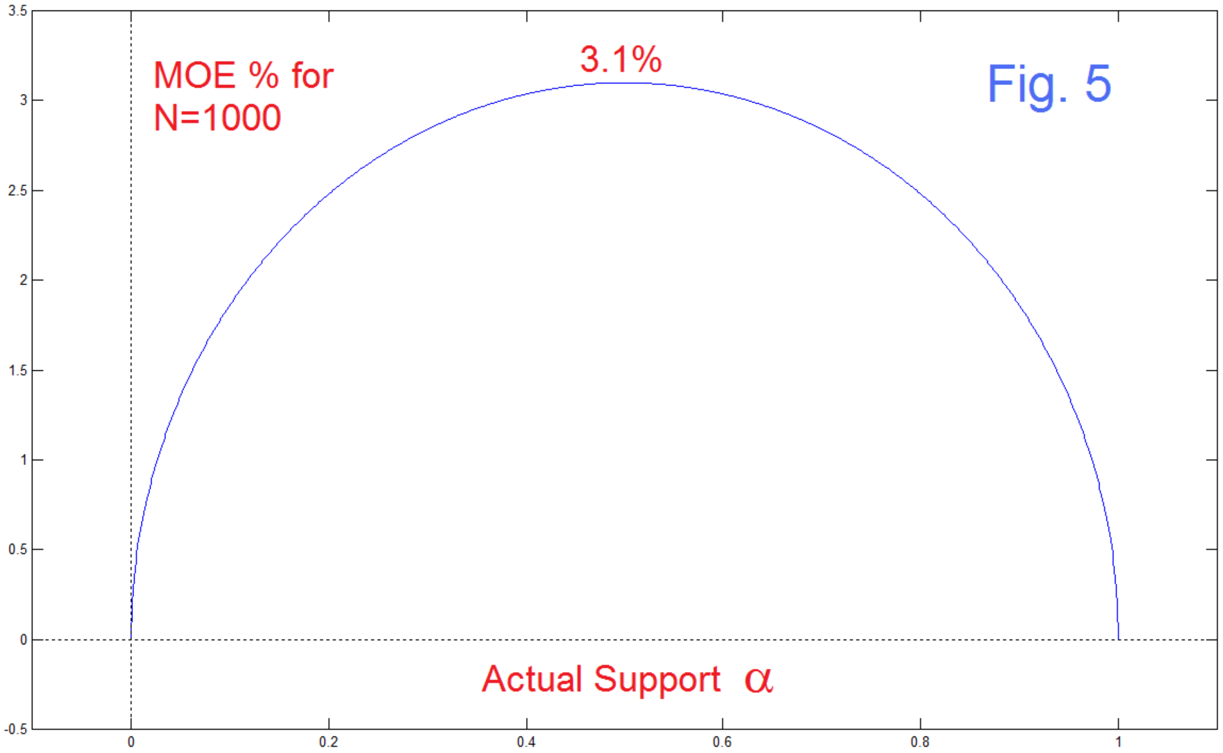


Fig. 5

3.5
3
2.5
2
1.5
1
0.5
0
-0.5
0
0.2
0.4
0.6
0.8
1
MOE % for
N=1000
3.1%
Actual Support α

These results are added to Fig. 4. and for comparison, a normal curve for the same μ and σ is shown in blue, using equation (3a). The height of the blue curve is arbitrary, but note that the extent exceeds the 0-100% limits of our supposed election poll. (Of course, any normal distribution tapers but never technically goes to zero.) [The notion of a negative vote is appealing – it would perhaps be appropriately humbling if the winning candidate was the one with the lesser negative score!]

If we try different values or actual support, we find that there is a MOE of almost exactly 3.1% (3.099%) at 50% support. We saw that at 40%, it was down to 3.04% and clearly it should taper to zero at both 0% and at 100% support. This is because as the race becomes more one-sided the more information the poll gives sooner. In fact, if the reporter says it correctly, they should say "maximum sampling error of about 3 percentage points". What is the function exactly?

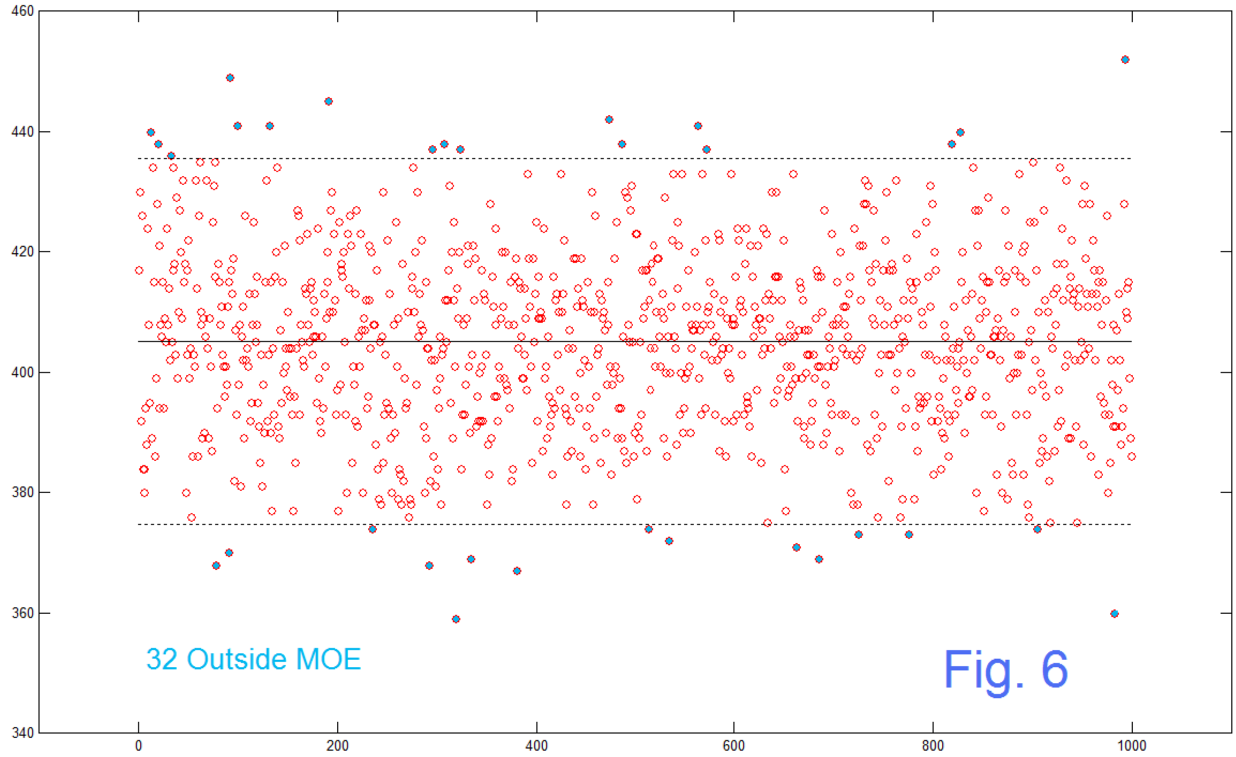
Say a candidate has the support of a fraction α of voters, and that there are only two candidates and no apathetic voters, so the opponent has the support of the fraction (1-α). The candidate thus receives a total of N×α×100% percentage points (all N×α give the candidate 100%). Of course the candidate also gets N×(1-α)×0% which adds nothing. The mean is N×α×100%/N or α×100% of course. Because the support is either 0% or 100%, and the mean is α×100%, all N points have errors. There are only two different errors 100(1-α) for αN of the points (those that support) and 100α for the (1-α)N remaining voters (who don't support). We square these, weight them by the corresponding number of voters, add them and divide by N, and take the square root [just equation (2)].

$$RMS = \sqrt{\frac{(1-\alpha)N(\alpha^2\ 100\%^2\ ) + \ \alpha N(1-\alpha)^2 100\%^2}{N}}\ = \sqrt{\alpha(1-\alpha)}\ 100\% \qquad (5)$$

and to find the MOE using equation (4) we have

$$MOE = \left[1.96\frac{\sqrt{\alpha(1-\alpha)}}{\sqrt{N}}\right] \qquad (6)$$

Equation (6) is plotted in Fig. 5 and note it is maximum at α=0.5 where it becomes very close to $1/\sqrt{N}$ which is here actually $0.98/\sqrt{1000}$ = 0.98/31.623 = 3.099%, as previously stated. Possibly we concentrate on MOE values of about 3% where we have two level polling because we are mainly concerned with issues that are closely contested so that α is in the center range around 0.5.



In the case of the speed data we found it useful to consider a theoretical case where we can simulate many many trials, which resulted in Fig. 3. We can do the same thing for the two candidate case, and this is shown in Fig. 6. Here we started with a population of 10,000, 40% of whom are assigned a random value 1, the remainder being assigned 0. We then sample these 1000 at a time. This we can do by starting at a random point in the 10,000 binary choices (say at an index from 0 to 50, randomly selected) and jumping through the data in steps (say 1 to 10, randomly selected). When we complete the 1000 points in each sample, we compute the mean, and the means of the successive samples are plotted in Fig. 6. We note a rather excellent result where 32 of 1000 (3.2%) are outside the MOE of 3.04. Ten additional runs of the same program that gave Fig. 6 yielded 25, 44, 35, 31, 29, 27, 32, 33, 36, and 32 (average 32.4) outside.

SOME END MATTERS

The most important thing to mention before quitting here is that the numbers you get doing this sort of calculation may not mean very much. In particular, the MOE kind of says you have taken enough samples (or not) assuming you did everything else right and your assumptions are valid. It is very easy to confound your results. So being inside the MOE doesn't prove a good deal. Better there than not.

Is it good (or good enough) to have your results inside the MOE 95% of the time? In fact, it seems fine when you are considering one result. That is a 5% (1 in 20) chance of making a mistake. Actually, that's pretty high if you keep right on running tests. The probability of getting things right is (19/20) and so if you run 20 experiments for example, the probability that they are all right is: just $(19/20)^{20}$ or 36%.

All of these issues are full of terminology, of which only the term "mean" (better, the average) is likely to resonate with the public. Things like standard deviation, margin-of-error, P-values, statistically-significant, confidence levels, and confidence interval, are all easily confused.

As this relates to "significance" we end by noticing that significance relative to statistics is not the same thing as significance as something you should care about or worry about. Using the traffic example, it might be true that two different surveys of speed give 50 mph and 52 mph and that the result is (nominally at least) statistically significant. That is, there is in some sense a real difference. But probably this difference would not be significant in the every-day sense. You might well say 50 or 52 – who cares! In the everyday sense though, 45 mph vs 60 mph would likely be something to worry about.

CODE USED

The Matlab code below is just for reference. It is all simple code just for making the pictures here, so is ordered as the figures in the note are 1 – 6. The corresponding Matlab figures are not in general the same. The code however may be useful as to showing exactly what was done – especially as there might be miscomprehensions in the presentation. Two other "snippets" of code used appear in the text.

FIGURE 1 of Note

```
clear
% make and check distributions centered at 50 and 55 sd=15
SD=8
SA=50
SB=55
A = 50+SD*randn(1,10000);
B = 55+SD*randn(1,10000);
% find beyond 2 sd for A
T=0;
```

```
for k=1:10000
   if A(k)<(SA-2*SD)
      T=T+1;
   end
   if A(k)>(SA+2*SD)
      T=T+1;
   end
   end
   T2=T/10000;
% and for limit of 1.96 sd
   T=0;
for k=1:10000
   if A(k)<(SA-1.96*SD)
      T=T+1;
   end
   if A(k)>(SA+1.96*SD)
      T=T+1;
   end
   end
   T196=T/10000;
%
% plot A and B data histograms
X=0:120;
HA=hist(A,X);
HB=hist(B,X);
MA=mean(A)
MB=mean(B)
STDA=std(A)
STDB=std(B)
figure(1)
plot(HA,'bo')
hold on
plot(HB,'ro')
plot([0 0],[-100 1000],':k')
plot([-10 130],[0 0],':k')
plot([MA MA],[-100 1000],'b')
plot([MB MB],[-100 1000],'r')
plot([MA-STDA MA-STDA],[-100 1000],'b:')
plot([MA+STDA MA+STDA],[-100 1000],'b:')
plot([MB-STDB MB-STDB],[-100 1000],'r:')
plot([MB+STDB MB+STDB],[-100 1000],'r:')
plot([MA-2*STDA MA-2*STDA],[-100 1000],'c:')
plot([MA+2*STDA MA+2*STDA],[-100 1000],'c:')
axis([20 90 -30 600])
hold off
figure(1)
T2
T196
```

FIGURE 2 of Note

```
figure(3)
t=-4:.01:4;
x=(1/sqrt(2*pi))*exp(-t.^2/2);
plot(t,x)
hold on
plot([-10 10],[0 0],'k')
plot([0 0],[-1 2],':k')
plot([-1 -1],[-1 2],'c')
plot([1 1],[-1 2],'c')
plot([-2 -2],[-1 2],'c')
plot([2 2],[-1 2],'c')
axis([-4 4 -.05 .5])
figure(3)
```

FIGURE 3 of Note

```
% Now take a large set of samples of length 1000 and look at means
C=1.96*SD/sqrt(1000)
NOp=0
NOm=0
AAM=zeros(1,10000);

for m=1:10000
   AA = 50+SD*randn(1,1000);
   AAM(m)=mean(AA);
   if AAM(m)>50+C
      NOp=NOp+1;
   end
   if AAM(m)<50-C
      NOm=NOm+1;
   end
end

figure(2)
NOp
NOm
NO=NOp+NOm
NOPERCENT=NO/10000
plot([0:1000],AAM(1:1001),'ro')
hold on
plot([-100 1100],[50+C 50+C],'k:')
plot([-100 1100],[50-C 50-C],'k:')
hold off
axis([-10 1010 49 51])
figure(2)
```

FIGURE 4 of Note

```
figure(1)
plot([0 0],[-100 1000],'k:')
hold on
plot([-150 150],[0 0],'k:')
plot([0 0],[0 600],'r')
plot(0,600,'ro')
plot([100 100],[0 400],'r')
plot(100,400,'ro')
plot([40 40],[0 1000],'b:')
%
x=[-150:150];
y=( 40000 / (49*sqrt(2*pi)) )* exp( - ( ( (x-40).^2 ) / (2*(49^2))    ) );
plot(x,y,'c')
plot([40-49 40+49],[450 450],'g')
hold off
axis([-150 150 -50 700])
```

FIGURE 5 of Note

```
alpha=0:.001:1;
RMS=sqrt(alpha.*(1-alpha))*100;
figure(1)
plot(alpha,RMS)
MOE=1.96*RMS/sqrt(1000);
figure(2)
plot(alpha,MOE)
hold on
plot([0 0],[-1 5],'k:')
plot([-1 2],[0 0],'k:')
axis([-0.1 1.1 -0.5 3.5])
hold off
figure(2)
MOE(490:510)
```

FIGURE 6 of Note

```
d= .4
s=rand(1,10000);
ss=zeros(1,10000);
for k=1:10000
   if s(k)<d
      ss(k)=1;
   end
end
SUM=sum(ss)
MEAN=mean(ss)
SD=std(ss)
MOE=1.96*SD/sqrt(1000)
%
for m=1:1000
    idx=ceil(50*rand);
    sss=zeros(1,1000);
   for k=1:1000
      sss(k)=ss(idx);
      idx=idx+ceil(10*rand);
   end
   tri(m)=sum(sss);
end
%
mtri=mean(tri)
%
figure(1)
plot([0:999],tri,'or')
hold on
plot([0 1000],[mtri mtri],'k')
plot([0 1000],[ mtri-30.4 mtri-30.4],'k:')
plot([0 1000],[ mtri+30.4 mtri+30.4],'k:')
hold off
axis([-100 1100 340 460])
figure(1)
```