A Better Type of Sample Mean?

by D Williams

Abstract:

A new class of sample means are examined which seem to better estimate population means for at least some distributions.

The standard arithmetic sample mean

$$\overline{X}_n = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

is NOT the best estimator of the population mean μ for all distributions.

For some distributions, there is a class of *ordered* sample means that provide a better estimate, more times than not.

The simplest such *ordered* sample mean is for sample size 3:

$$\overline{X}_{alt(3)} = \frac{3x_1 + 2x_2 + 3x_3}{8} \quad where\ x_1 \le x_2 \le x_3$$

- where x values have been swapped so that
  $x_1 = \min\{x_1,x_2,x_3\}$
  $x_2 = \text{median}\{x_1,x_2,x_3\}$
  $x_3 = \max\{x_1,x_2,x_3\}$
  to provide an ordered sample.

This is often a better estimator of μ than the normal sample mean.

For instance, consider this very simple example:

Draw a numbered ball 3 times with replacement from 2 balls numbered α and β with α ≤ β. Which is the better sample mean?

There are 8 arrangements of 3 selections from 2 choices with 4 ordered possibilities.

Draw (ordered)	*# of*	*Alt sample mean*	*Standard sample mean*	*Absolute difference of alt-SM from μ*	*Absolute difference of alt-SM from μ*	*Alt-SM closer?*
(α,α,α)	1	α	α	(β-α)/2	(β-α)/2	same
(α,α,β)	3	(5α+3β)/8	(2α+β)/3	(β-α)/8	(β-α)/6	yes
(α,β,β)	3	(3α+5β)/8	(α+2β)/3	(β-α)/8	(β-α)/6	yes
(β,β,β)	1	β	β	(β-α)/2	(β-α)/2	same

The alt-sample mean is closer to the population mean (=(α+β)/2) 6 out of 8

times, with the other 2 being the same distance from μ.

Other continuous distributions so tested seemed to give a similar result.

For instance take the stochastic function f(r)=(rand#)^(1/3) - equivalent to $pr(x)=3x^2$ for $0 \leq x \leq 1$ - from which 3 samples are taken.

To test this I used a spreadsheet with 125 points evenly spread between 0 and 1 with f(x)=x^(1/3).

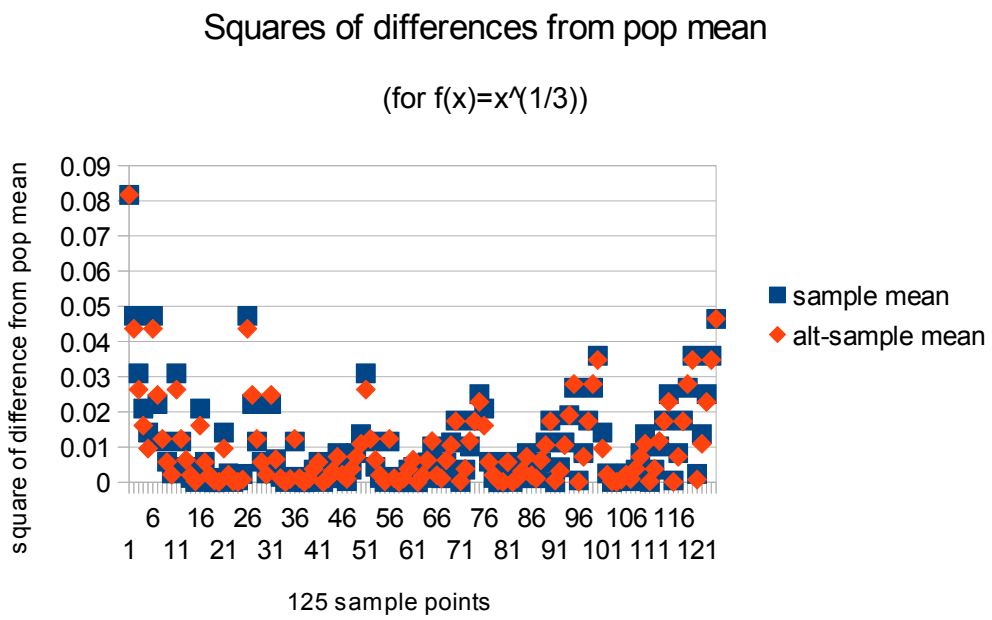
The random numbers were "represented" by triplets from (0.1,0.1,0.1) to (0.9,0.9,0.9) in steps of 0.2 across all 125 possible arrangements (from 5 possible choices for first value – 0.1,0.3,0.5,0.7,0.9 and similar for the second and third values – giving 125 triplets).

I then compared the standard sample mean with the new one.

Calculating the sum of the squares of the differences from the population mean (here ¾) gave the following values:

Sample mean: 1.29946
Alternative sample mean: 1.24116

Notice that the alternative gives the closer result.



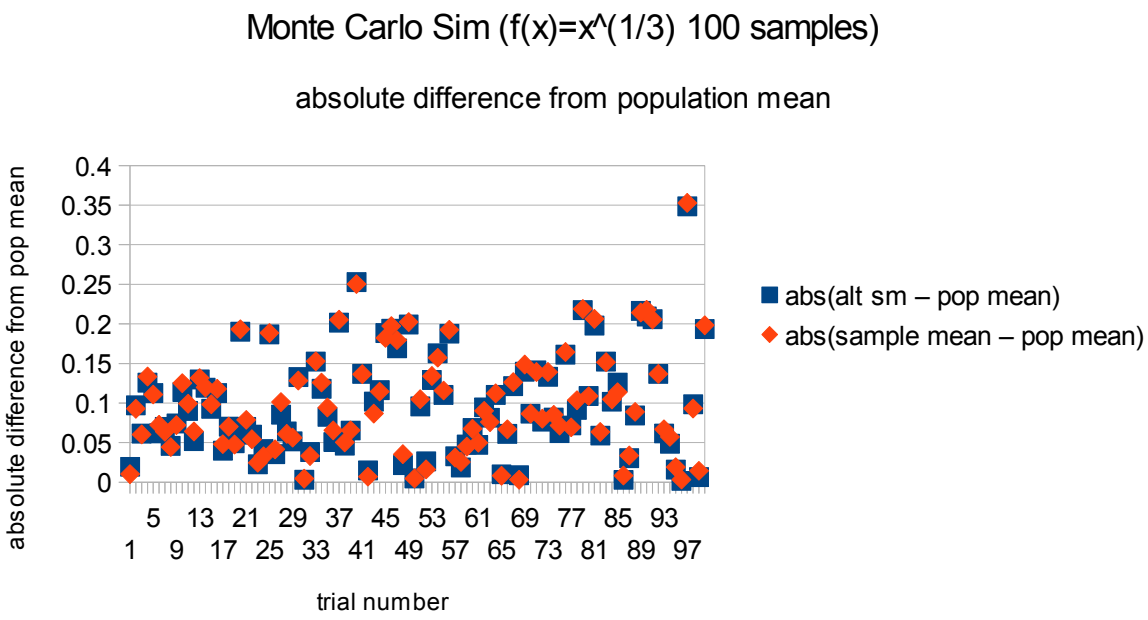
Graph: *result of spreadsheet test using 125 evenly spaced random "surrogates"*

In 72 cases (out of 125) the alt sample mean was closer to the population mean than the regular sample mean and was the same 5 times.

Thus, in this example, the values *suggest* that the new sample mean is better at estimating the population mean than the standard sample mean.

Monte Carlo tests back up this result.

In the plot of one Monte Carlo simulation below, in 63 times out of 100 the alt sample mean is closer to the population mean than the regular sample mean.



Which distributions (stochastic functions) display this improved result for the alt sample mean? I have no idea. Hopefully readers can help examine this matter further.

Higher order alt sample means exist.

The *ordered* sample mean for sample size 4 is:

$$\overline{X}_{alt(4)} = \frac{13x_1 + 11x_2 + 11x_3 + 13x_4}{48} \quad where\ x_1 \le x_2 \le x_3 \le x_4$$

- again after sample values x have been swapped to produce an ordered sample.

As per the n=3 example above, drawing 4 balls with replacement from 2 balls numbered α and β with α ≤ β gives (after ordering) the following estimates by sample mean.

Draw (ordered)	***# of***	***Alt sample mean***	***Standard sample mean***	***Absolute difference of alt-SM from μ***	***Absolute difference of alt-SM from μ***	***Alt-SM closer?***
(α,α,α,α)	1	α	α	(β-α)/2	(β-α)/2	same
(α,α,α,β)	4	(35α+13β)/48	(3α+β)/4	11(β-α)/48	12(β-α)/48	yes
(α,α,β,β)	6	(24α+24β)/48	(2α+2β)/4	0	0	same
(α,β,β,β)	4	(13α+35β)/48	(α+3β)/4	11(β-α)/48	12(β-α)/48	yes
(β,β,β,β)	1	β	β	(β-α)/2	(β-α)/2	same

Thus the alt-mean is closer 8 times out of 16 and the same 8 times.

Here is a small test program approximating the alt sample mean size 4 over a "representative" spread of samples for f(x)=x^(1/3).

```
popmean=3/4

ctrless=o: ctrsame=0: ctr=0

for i=0.1 to 0.9 step 0.2
 for j=0.1 to 0.9 step 0.2
  for k=0.1 to 0.9 step 0.2
   for l=0.1 to 0.9 step 0.2

a=i^(1/3)
b=j^(1/3)
c=k^(1/3)
d=l^(1/3)

maxterm=max(a,b,c,d)
minterm=min(a,b,c,d)

altsm=(13/48)*(minterm+maxterm)+(11/48)*(a+b+c+d-minterm-maxterm)

sm=(a+b+c+d)/4

if abs(altsm-popmean)<abs(sm-popmean) then ctrless=ctrless+1

if abs(altsm-popmean)=abs(sm-popmean) then ctrsame=ctrsame+1

ctr=ctr+1

next l
next k
next j
next i

print ctr,ctrsame,ctrless

stop
```

The free software (SmallBasic) was used to run the program.

When run, it shows that the alt sample mean is closer to the population mean than the standard sample mean 373 times out of 625, is the same 52 times, and further away 200 times. That is, it's better most of the time.

When using the above program, suitably modified for different stochastic functions, the following results were obtained for 6 arbitrarily chosen functions:

Stochastic function	*Alt-sample mean closer?*	*Alt-sample mean same?*	*Alt-sample mean further?*
2x^3+7x+3	381	58	186
x^(1/3)	373	52	200
10sin(pi*x/2)	372	65	188
10tan(pi*x/4)	368	65	192
exp(x)	380	65	180
ln(1+x)	380	59	186

In each case, the alt-sample mean is closer to the population mean μ than the standard sample mean more often. Is this always the case?

As a further test, a Monte Carlo Simulation was undertaken using 6 stochastic functions using the program shown in the endnotes. It gave the following results:

Monte Carlo Simulation (1000 samples each simulation):

Stochastic function	*Alt-sample mean closer than normal sample mean?*	*Alt-sample mean further away than normal sample mean?*
2x^3+7x+3	649	351
x^(1/3)	559	441
10sin(pi*x/2)	651	349
10tan(pi*x/4)	653	347
exp(x)	658	342
ln(1+x)	627	373

Repeated simulations (5 trials of 1000 samples for the above 6 stochastic functions) produced similar results. In each case, the alt-sample mean was more often closer to the population mean μ than the standard sample mean.

The *ordered* sample mean for sample size 5 is:

$$\overline{X}_{alt(5)} = \frac{275}{1152}x_1 + \frac{25}{288}x_2 + \frac{67}{192}x_3 + \frac{25}{288}x_4 + \frac{275}{1152}x_5$$
$$where\ x_1 \le x_2 \le ... \le x_5$$

The *ordered* sample mean for sample size 6 is:

$$\overline{X}_{alt(6)} = \frac{247}{1280}x_1 + \frac{139}{1280}x_2 + \frac{127}{640}x_3 + \frac{127}{640}x_4 + \frac{139}{1280}x_5 + \frac{247}{1280}x_6$$
$$where\ x_1 \le x_2 \le ... \le x_6$$

These equations for the alt sample mean come from an alternative model of

probability theory previously developed (see "An Alternative Model of Probability Theory" at vixra.org/author/d_williams). Under this alternative version, the Central Limit Theorem looks like:

$$\Pr\left(\frac{\frac{S_n}{n} - \int_0^1 f(x)dx}{\frac{\sigma}{\sqrt{n}}} < \beta \right) \to \Phi(\beta)$$
$$as\ n \to \infty\ for\ suitable\ f(x)$$

Now that the population mean has the form of an area, the standard arithmetic sample mean can be likened to the Mid-Point Rule approximation. Thus better approximations of the "area" of the population mean should produce a better sample mean.

Unfortunately I could not use Simpson's Rule or the Trapezoidal Rule due to the x-values they use (somebody might be able to find a fix around this?).

What I did was to order the sample of n values ($y_1 \le y_2 \le ... \le y_n$), assign the f-values as $f(1/2n)= y_1$, $f(1/2n)= y_2$, … , $f((2n-1)/2n)= y_n$ then calculate an n-th order polynomial through the n points by solving the appropriate n simultaneous equations.

For example, with n=3, sample then order $y_1 \le y_2 \le y_3$ by swapping values then assign $f(1/6)=y_1$, $f(3/6)=y_2$, $f(5/6)=y_3$

Now apply a quadratic curve ax^2+bx+c through the data points $(1/6,y_1)$, $(3/6,y_2)$, $(5/6,y_3)$ then solve the 3 simultaneous equations

$a(1/6)^2+b(1/6)+c=y_1$
$a(3/6)^2+b(3/6)+c=y_2$
$a(5/6)^2+b(5/6)+c=y_3$

This gives the quadratic equation

$ax^2+bx+c=(9/2)(y_3-2y_2-y_1)x^2+3(3y_2-y_3-2y_1)x+(1/8)(15y_1-10y_2+3y_3)$

Now take the integral of this from 0 to 1 to give the alternative sample mean of

$(3/8)y_1+(1/4)y_2+(3/8)y_3$

The same method is used for higher order versions
(For n=1 or 2, you just get the standard sample mean).

In general, the n-th alt sample mean is the integral

$$\int_0^1 (ax^{n-1} + bx^{n-2} + cx^{n-3} + ...)dx$$

where

$$\begin{bmatrix} a \\ b \\ c \\ \dots \end{bmatrix} = \begin{bmatrix} \left(\frac{1}{2n}\right)^{n-1} & \left(\frac{1}{2n}\right)^{n-2} & \dots & 1 \\ \left(\frac{3}{2n}\right)^{n-1} & \left(\frac{3}{2n}\right)^{n-2} & \dots & 1 \\ \dots & \dots & \dots & 1 \\ \left(\frac{2n-1}{2n}\right)^{n-1} & \left(\frac{2n-1}{2n}\right)^{n-2} & \dots & 1 \end{bmatrix}^{-1} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$$

where x_i are the n samples that have been ordered such that $x_1 \leq x_2 \leq \ldots \leq x_n$.

See below for more details on constructing such matrices.

There are still lots of questions:

1. What do higher order alt sample means look like? Do they work?
2. What distributions exhibit the benefits outlined above?
3. Are there other, possibly better, types of sample means (not necessarily arithmetic)?
4. Can other types of approximate quadrature be adapted to make other types of sample and population means? (Simpson's Rule? Gaussian quadrature?)
5. Can new versions of population means be devised?
6. Can simpler ways of calculating alt sample means be devised?
7. Can other models of probability theory be developed and similarly used?
8. Are there real world examples where the alt means provide benefit?

For example, with question 4, it is possible to join alt sample means of size 3 together for samples of size 5, 7, 9, etc to construct another alt sample mean. Whether it is better or worse, I haven't looked at yet.

Other types of means like geometric means also need looking at.

For instance is

$$\overline{X}_{alt-geo(3)} = (x_1 x_3)^{\frac{3}{8}} (x_2)^{\frac{2}{8}} \; ?$$
$$where \; x_1 \leq x_2 \leq x_3$$

a better geometric sample mean than the standard one?

For question 5, with the f(x) formulation of probability theory, it is possible to replace the population mean with a sequence of "partial" population means staring with f(1/2) for sample size 1, (f(1/4)+f(3/4))/2 for size 2, (f(1/6)+f(1/2)+f(5/6))/3 for size 3 and so on for non-decreasing f(x). This only works if you know f(x), of course.

So there is a lot to do.

Someone might also provide some simple-to-use freely accessible code for others to make investigations easier (not my speciality).

Endnotes

1. Program for Monte Carlo simulation

(1000 samples, f(x)=2x^3+7x+3 used as example)

```
        popmean = 7

        ctrless=0:ctrbigger=0

        altsm=0:sm=0

for i=1 to 1000

        a=RND
        b=RND
        c=RND
        d=RND

 X=2*(a^3)+7*a+3
 Y=2*(b^3)+7*b+3
 Z=2*(c^3)+7*c+3
 W=2*(d^3)+7*d+3

                maxterm=max(X,Y,Z,W)
                minterm=min(X,Y,Z,W)

        sm=(X+Y+Z+W)/4

        altsm=(13*(maxterm+minterm)+11*(X+Y+Z+W-maxterm-minterm))/48

  if abs(altsm-popmean)<abs(sm-popmean) then ctrless=ctrless+1

     if abs(altsm-popmean)>abs(sm-popmean) then ctrbigger=ctrbigger+1


 next i
```

```
        print ctrless, ctrbigger

stop
```

2. Constructing inverse matrices used above

The free software wxMaxima was used to invert the relevant matrices below:

For the alt sample mean of size 3, the relevant matrix is:

```
matrix(
[(1/6)^^2, (1/6), 1],
[(3/6)^^2, (3/6), 1],
[(5/6)^^2, (5/6), 1]
);
```

that is:

$$\begin{bmatrix} \frac{1}{36} & \frac{1}{6} & 1 \\ \frac{1}{4} & \frac{1}{2} & 1 \\ \frac{25}{36} & \frac{5}{6} & 1 \end{bmatrix}$$

which inverted gives:

$$\begin{bmatrix} \frac{9}{2} & -9 & \frac{9}{2} \\ -6 & 9 & -3 \\ \frac{15}{8} & -\frac{5}{4} & \frac{3}{8} \end{bmatrix}$$

For the alt sample mean of size 4:

```
matrix(
[(1/8)^^3, (1/8)^^2, 1/8, 1],
[(3/8)^^3, (3/8)^^2, 3/8, 1],
[(5/8)^^3, (5/8)^^2, 5/8, 1],
[(7/8)^^3, (7/8)^^2, 7/8, 1]
);
```

that is:

$$\begin{bmatrix} \frac{1}{512} & \frac{1}{64} & \frac{1}{8} & 1 \\ \frac{27}{512} & \frac{9}{64} & \frac{3}{8} & 1 \\ \frac{125}{512} & \frac{25}{64} & \frac{5}{8} & 1 \\ \frac{343}{512} & \frac{49}{64} & \frac{7}{8} & 1 \end{bmatrix}$$

when inverted gives:

$$\begin{bmatrix} -\frac{32}{3} & 32 & -32 & \frac{32}{3} \\ 20 & -52 & 44 & -12 \\ -\frac{71}{6} & \frac{47}{2} & -\frac{31}{2} & \frac{23}{6} \\ \frac{35}{16} & -\frac{35}{16} & \frac{21}{16} & -\frac{5}{16} \end{bmatrix}$$

and so on.

It would be interesting to know what properties these inverse matrices have.

3. Outline of Alternative Version of Probability Theory

Standard Probability Theory	***Alternative Probability Theory***
$\mu = E(x) = \int_{-\infty}^{\infty} x * pr(x)dx$	$\mu = E(x) = \int_0^1 f(x)dx$ $where\ f(x) = inv(\int_{-\infty}^{x} pr(x)dx)$
$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2\, pr(x)dx$ $= E(x^2) - [E(x)]^2$	$\sigma^2 = \int_0^1 (f(x))^2 dx - (\int_0^1 f(x)dx)^2$
$\Pr\left(\frac{\frac{S_n}{n} - \int_{-\infty}^{\infty} x * pr(x)dx}{\frac{\sigma}{\sqrt{n}}} < \beta \right) \to \Phi(\beta)$ $as\ n \to \infty$	$\Pr\left(\frac{\frac{S_n}{n} - \int_0^1 f(x)dx}{\frac{\sigma}{\sqrt{n}}} < \beta \right) \to \Phi(\beta)$ $as\ n \to \infty$

$where\ S_n = x_1 + x_2 + ... + x_n$

$and\ inv = the\ inverse\ operation$