# In the Beginning of Statistics

A Two Act Play<br>
By Patrick Staley

Act One: The First Puzzle

Narrator:
One day God decided to test Man. He said to Man,

God:
I have three numbers and I would like you to guess the best single number to represent these three numbers.

Narrator:
In the scriptures these three numbers are called the "Population".

But the mathematically inclined among the humans contested,

Math Rabble:
But God how shall we know what means the word "best"?

Narrator:
And God replied,

God:
I shall judge your answers by computing the square of the difference between your guess and each of the three numbers. And this value shall be called the "squared error". Whomsoever gets a low squared error will have pleased me.

[Tablet in the sky with the following:
$(x_i - guess)^2$]

Narrator:
But to this the humans were still confused. They pleaded with God saying,

Humans:
But God, since the Population has three numbers and we have only one guess, how shall we decide among the three squared errors, which is the most important?

[Tablet in the sky with

$(x_1 - guess)^2, (x_2 - guess)^2, or\,(x_3 - guess)^2$ ?]

Narrator:
And God responded,

God:
All three squared errors are equally important. These squared errors shall be summed and we shall call the result the sum of the squared error. Whomsoever gets a sum of the squared errors as low as God's is successful.

[Tablet in the sky
minimize $(x_1 - guess)^2 + (x_2 - guess)^2 + (x_3 - guess)^2$
]

Narrator:
And this procedure for measuring "best" came to be called the method of the "Least Squares".

Now among the humans there were disciples of Newton and Leibnitz whose way was called the Calculus. These followers of Newton and Leibnitz ministered to God's best guess problem as follows:

Calculati:
Let the three numbers be called $x_1$, $x_2$, and $x_3$.
Let our guess be called x.
Then the Sum of the Squared Errors (SSE) will be

[At this point the message appears on a tablet above the stage
$SSE = (x_1 - x)^2 + (x_2 - x)^2 + (x_3 - x)^2$
]

From the principles of our great teacher, Newton, we know that the minimum for this expression will be given when the derivative is zero.

Narrator:
At this point those who were not of the Calculus sect and knew not the ways of the Calculati became confused. The confused ones found their own solution which they called "Finding the vertex of a parabola".

[Separate tablet appears with
Vertex of $y = ax^2 + bx + c$ is at $x = -b/(2a)$
a picture also]

Now the Calculati proceeded as follows,

Calculati:
The derivative of SSE is

[Now the magic tablet showed this expression
$-2(x_1 - x) - 2(x_2 - x) - 2(x_3 - x)$
]

We can set this expression to zero,

[Tablet switches to
$0 = -2(x_1 - x) - 2(x_2 - x) - 2(x_3 - x)$
]
solve the resulting equation, and know the answer to God's puzzle.

Narrator:
And thus the Calculati came to know the answer to be $(x_1+x_2+x_3)/3$.
This number came to be called the Population Mean and
it was henceforth known by the Greek letter $\mu$.

And it came to pass that the solution called "Finding the vertex of a parabola" yielded the same answer to God's puzzle.

Now the humans were pleased. They knew how to solve God's Puzzle. So they came unto God and said

People:
God we are ready to be tested. Tell us your three numbers and we shall give you the best representative for those numbers.

Narrator:
Now God said to himself,

God:
These humans are clever. I shall teach them the lesson of living with uncertainty.

Narrator:
Then God said to the humans,

God:
Pick three among you to receive numbers from me. You shall call these three the three Scientists.

Narrator:
And the humans, eager to please God, presented him with three scientists and they were called Scientist One, Scientist Two, and Scientist Three.

Now God said unto the Three Scientists,

God:
To each of you I will give exactly two numbers from the Population. You shall call your two numbers Sample One, Sample Two, or Sample Three to match you own names. Each of the samples shall be different. You must each answer my puzzle separately without knowing all three of the numbers. Nor shall you know the guesses of the other Scientists.

Narrator:
And once again the humans were thrown into great confusion. For they knew how to solve God's Puzzle knowing all three numbers of the Population but God had forbidden their scientists from knowing the entire Population—the Scientists were to know only that part of the Population that was their own Sample.

So, much humbled, the humans returned to God with this plea,

Humans:
God, tell us how our scientists shall be judged so that we may make our representative guesses good ones.

Narrator:
Then God said to the humans,

God:
The Scientists shall be judged collectively. I see that you have understood the principle that the Mean is the representative that minimizes the Sum of the Squared Errors. This is good. For your three guesses I shall divine the mean of these three guesses and that shall be the collective guess of humanity.

Narrator:
Now there was great confusion among the humans. Finally the humans decided that there was nothing better to do than to follow the example of God and so the first scientist, whose sample was {$x_1$, $x_2$}, computed the mean from his sample—

$$\overline{x_1} = \frac{x_1 + x_2}{2}$$

and submitted this as his guess.

Scientists Two and Three did likewise. Submitting their guesses as

$$\overline{x_2} = \frac{x_2 + x_3}{2}$$

and

$$\overline{x_3} = \frac{x_1 + x_3}{2}$$

These three numbers came to be known as the Sample Means.

Lo and Behold God praised the Scientists and said that their procedure was good. And god sent Al-Khwarizmi to the humans with the gift of Elementary Algebra. And with this gift the Algebraists among the humans were able to mimic God. That is they used the test God had specified for evaluating the collective guess of the humans.

Now this meant that they would compare the mean of the three Sample Means

$\overline{x_1}, \overline{x_2}, and\ \overline{x_3}$

with the Population mean

$$\frac{x_1 + x_2 + x_3}{3}$$

They reasoned as follows:

$$\frac{\overline{x_1} + \overline{x_2} + \overline{x_3}}{3} = \frac{\frac{x_1 + x_2}{2} + \frac{x_2 + x_3}{2} + \frac{x_1 + x_3}{2}}{3}$$

$$= \frac{\frac{2x_1 + 2x_2 + 2x_3}{2}}{3}$$

$$= \frac{x_1 + x_2 + x_3}{3} = \mu$$

And God saw that man understood his great lesson. And this lesson came to be known as

## The Mean of the Sample Means is the Population Mean

End of Act one

## Act Two: The Mystery of n Minus One

[Note: In the interest of Political Correctness God has decided to change gender for Act Two]

Narrator:
After many ages God saw that Man was complacent. And God was displeased. So God set about to humble man. She called forth all those who had learned the lesson of the Mean of the Sample Means. And She said:

God:
It is good that you have learned to compute the mean and to use it as the best representative by the rule of the Least Squares. This lesson you have learned well and I am pleased.

But now I challenge you to tell me the value of this Least Squared Error.

Narrator:
Now the people were confident because they knew how to compute the mean. Further more they understood what God wanted them to tell Her. And they could compute the answer using the formula--

[Now the magic tablet showed this expression]
$SSE = (x_1-\mu)^2 + (x_2-\mu)^2 + (x_3-\mu)^2$ . . . until the end of the Population

So the people replied to God,

People:
Tell us the Population and we shall answer God's challenge.

Narrator:
Now God knew well the ways of Man. Man had eaten from the fruit of the Least Squared Procedure Tree. God's plan was to test the Scientists with the task of Inference from a Sample. And thus God spoke unto the People,

God :
I have again a Population of three numbers. Bring from among you three Scientists who shall guess the Sum of the Squared Errors of the Population Mean. You shall call the three Scientists Scientist One, Scientist Two and Scientist Three.

Narrator:
Once again the people selected three among them to be the three Scientists. These three approached God. As in past times She challenged them with the uncertainty of knowing only a portion of the Population. She instructed them with these words,

God:
To each of you I will give exactly two numbers. You shall call your pair of numbers Sample One, Sample Two, or Sample Three to match your own names. Each of the samples shall be different. You must each answer my puzzle separately without knowing all three of the numbers. Nor shall you know the guesses of the other Scientists.

Narrator:
Now these Scientists knew of the exploits of the Scientists in Act One. So even though they were uncertain of their way, they decided to follow the way of the Scientists of Act One and mimic God's computation. Now God's computation of the SSE for the population was well known.
[sky tablet

$$\mu = \frac{x_1 + x_2 + x_3}{3}$$

$$SSE = (x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2$$

]

Thus Scientist One proposed estimating the Population SSE as
[sky tablet

$$\bar{x} = \frac{x_1 + x_2}{2}$$

$$SSE = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2$$

]

But Scientist Two objected to this procedure saying,

Scientist Two:
But God's SSE has three terms where as your estimate only has two terms. Clearly we should take the estimate from Scientist One's procedure and multiply by 3/2 to account for the difference in size between the Population and the Sample.

Narrator:
Now Scientist Three was uncertain which of the other two had the best answer. So Scientist Three proposed that they go back to God for help. Scientist Three expected that God would teach them the lesson of experimentation. So the three scientists approached God and Scientist Three appealed to God thusly,

Scientist Three:
We have studied the task you have assigned us. We are divided as to which is the best solution we should offer up to you. Perhaps you can offer us some practice tests so that we may know which of our methods is better?

Narrator:
To which God responded,

God:
For Heaven's sake, make up your own practice tests. What do you take me for, your math teacher?

Narrator:
Thus as God had rebuked the Scientists, they longed for the Eden of their Community College Math Classes. Their Community College Math Teachers had always made practice tests for them.

But after a while they decided to do as God had bid and make their own practice tests. They chose three numbers as a Population to make their own practice test. And the first practice test was {12, 12, 0}.

From this set they decided the three Samples would be {12,12}, {12,0}, and {12,0}. These sets begat the three Sample means
$\overline{x_1} = 12$
$\overline{x_2} = 6$
$\overline{x_3} = 6$

From this generation of answers was begat the three Sample SSE values
$SSE_1 = (12-12)^2 + (12-12)^2 = 0$
$SSE_2 = (12-6)^2 + (0-6)^2 = 72$
$SSE_3 = (12-6)^2 + (0-6)^2 = 72$

Now if they used the strategy of Scientist One they would submit the three SSE guesses as 0, 72 and 72.
Whereas if they used the multiply-by-3/2 strategy recommended by Scientist Two, they would submit the three guesses as 0, 108 and 108.

Scientist Three:
We shall see how God will grade these answers. Thus we shall know which is correct.

Narrator:
Their computations for God and the SSE of the Population proceeded as follows:

$$\mu = \frac{12+12+0}{3} = 8$$

$$SSE = (12-8)^2 + (12-8)^2 + (0-8)^2 = 96$$

Then God would compute the mean of their guesses as
(0 +72+72)/3 = 48 for the Scientist One strategy
and
(0+108+108)/3 = 72 for the Scientist Two strategy.

Now the Scientists saw that multiplying by 1.5 was better, but the result was still below God's target of 96.

.
So they returned to God for further guidance.

Scientist Three:
God, we have done as you said and made our own practice test. It appears that just computing the SSE for our samples yield a result that is too low. One among us believes this is because you have three in the Population whereas our Sample only has two. Yet when we correct for this by multiplying by 3/2 our answer is still too small.

Narrator:
At this point God had pity for the Scientists and helped them with these words,

God:
That the Population is bigger than the Sample and that you must correct for this is correct. If each of you knew the Population mean, $\mu$, and used that for your SSE computation then the method of Scientist Two would yield optimal results. However in computing the squared errors you used the Sample Means $\overline{x}_i$, rather than the Population Mean, $\mu$. When you use the sample mean you necessarily get smaller Squared errors because of the principle that the mean minimized the sum of the squared errors. Thus the method of multiplying by 3/2 will also yield an estimate of the SSE that is too small.

Narrator:
Thus were the three scientists confused and they left this problem for a long time.

After a while the Scientists decided to try other practice tests. In this way they hoped to pick a number bigger than 1.5 to scale by. These were the generations of their practice tests and the indicated scale factors:

| Population | | | Samples | | Scaling Required |
|---|---|---|---|---|---|
| | mean | SSE | SSE's | Mean SSE | |
| 0,6,12 | 6 | 72 | 18, 18, 72 | 36 | 2 |
| 0,9,12 | 12 | 378 | 40.5, 162, 364.5 | 189 | 2 |
| 1,2,3 | 2 | 2 | 0.5,0.5,2 | 1 | 2 |
| 1,7,10 | 6 | 42 | 18,4.5,40.5 | 21 | 2 |

Thus did the Scientists come to suspect that the correct scaling factor was two.

Now the Scientists returned to God confident that their answer would please God

Scientists:
We have proceeded with practice tests as you have bade us. From our practice tests we have decided to use twice the SSE of the Sample as our estimate for the SSE of the Population, which is known only to Her.

God:
You Scientists have done well. You have used the scientific principle of repeated experiments to achieve a convincing answer.

Narrator:
But then God went on,

God:
Lest anyone say that you were lucky I bade you to call upon the disciples of Al-Khwarizmi , those who know the ways of Elementary Algebra. Submit your proposal to them so that they may proclaim the gospel according to mathematics, which is universal and not subject to luck.

Narrator:
Now the scientists did as they were told and by and by the mavens of Elementary Algebra were able to establish that the correct scale factor was two, as the scientists had predicted. They presented the algebraic proof to God and She responded.

God:
What you have done is good. I implore you to post it on the internet so that others may learn the truth.

Narrator
And the proof was posted on the internet. The URL for the proof is
http://www.mr-ideahamster.com/stat-think/n-minus-one.pdf

## Epilogue

God went on to test man with populations of different sizes and samples of different sizes. In all of these cases the scriptures indicate that the procedures were the same. God would call for as many scientists as there were different samples of size n. The humans would run many practice tests. In all of these cases there was a specific scale factor that depended only on the size of the Population and the size of the Sample. And this scale factor yielded the best estimate of the Population SSE from the Sample SSE.

At God's request a dynamic means for doing these practice tests was incorporated into an Excel spreadsheet. This spread sheet was posted on the internet:

http://www.mr-ideahamster.com/stat-think/nm1study.xls

Here is a table with some values for the scale factor from the Excel worksheet:

| N | n | scale factor |
|---|---|---|
| 3 | 2 | 2/1 |
| 6 | 5 | 5/4 |
| 6 | 4 | 5/3 |
| 6 | 3 | 5/2 |
| 6 | 2 | 5/1 |

Notice that in all of these cases the scale factor is (N-1)/(n-1). This suggests that a helpful normalization would be to divide all SSE computations by one less than the number of terms. This value SSE/(n-1) is called the Variance. Thus for all sample sizes the Variance = SSE/(n-1) is the best estimate to the SSE/(N-1) for the Population or for the Variance for any other size sample. This estimate of variation is called unbiased because it is neither systematically too low nor is it too high.

The End