A Methodology to improve Goodness of Fit when Replicating Empirical Data utilizing Artificial Intelligence

Ahmad A. Moreb[1] & Naif Nahi Alharbi[2*]

King Abdul-Aziz University, Faculty of Engineering, Jeddah, Saudi Arabia,
E-mail: *prof.moreb@gmail.com*[1]

[*]Corresponding Author Email: *noon1985@live.com*[2]

***Abstract*:** Scientists and practitioners frequently resort to replicating empirical data when testing the validity of scientific theories or testing hypothesis. Commonly known probability distribution (Normal, Binomial, Exponential, etc.) are habitually assumed to fit the empirical data. To avoid complicated probability distributions, analysts find themselves tolerating poor values for the goodness of fit. In this paper, a methodology is introduced for replicating empirical data that succeeded in obtaining goodness of fit close to 100% compared to 87% goodness of fit using a known probability distribution. Moreover, Artificial Intelligence (AI) is developed spatially for this research to enhance accuracy further.

Keywords: Actual Distribution, Goodness of Fit, Artificial Intelligence, Integration.

I. INTRODUCTION

Many theoreticians and practitioners' resort to simulating the environment in which theories were based on, to be tested and validated. Replicating data in which the theories are based on is the first approach to be used in testing the hypothesis and presumption of their theories. When replicating data, it is customary to fit the empirical data to a commonly known distribution (i.e. Normal, Binomial, Exponential, etc.) intending to get the best fit for the empirical data [1-2]. In many cases, a good fit cannot be obtained; in which case, a poor fit is being tolerated to avoid complicated distribution requiring an expert statistician to handle it. Thus, ending up with conclusions that are far from accurate.

In this paper, we introduce a methodology that replicates the actual empirical data much closer to the real distribution. An example of real-life empirical data for 917 trainees (a BMI equation based on age, height, weight and fitness rate of each trainee) is drawn from the records of a Military Training Institute (MTI) [3]. The best common probability distribution was found to be a Normal Distribution, with a mean of (36.07) and a standard deviation of (8.35). Data covers the range of ages of men that extends from 20 years to 52 years old, as shown in "Fig. 1". Besides, they create a training plan based on that distribution and the percentage of each group to provide the training needs during the training year.

A Simulation is to be used to design future training plans. The goodness of fit for both, the replicated data using fitted normal probability distribution versus the replicated data using the proposed method.

 ©2020, IJA-ERA - All Rights Reserved

II. PROPOSED METHODOLOGY

A. *Choosing the correct index*

At first, we did not rely on age only in collecting and distributing data, but we also relied on the body mass of trainees (BMI) because it is the correct index in addition to age to design the appropriate training plan.

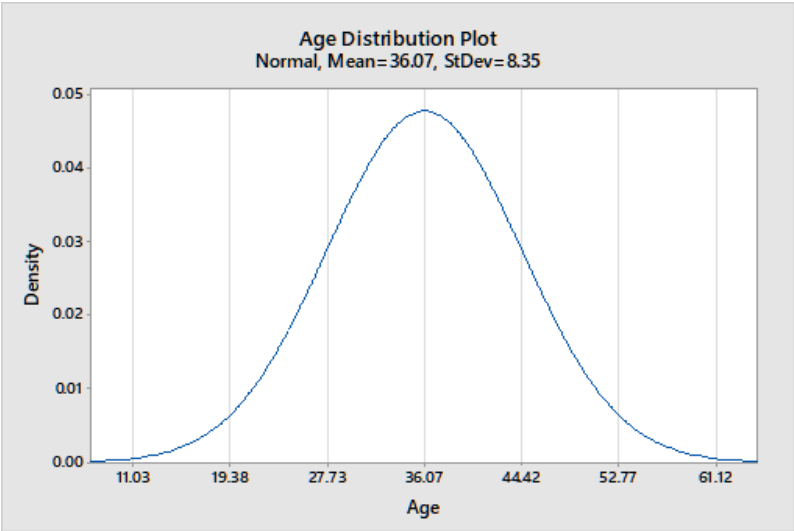

Figure 1: The MTI assumed distribution

Using the BMI equation based on the weight and length of the trainee, we get the body mass of each trainee through which we will divide them into categories by body mass, as in (1) [4].

$$BMI = \frac{weight\ (kg)}{Length^2\ (meter)} \quad (1)$$

BMI for trainees ranged from 15 to 55, we divided them into 8 categories to determine the number of trainees and Area in each category as shown in Table 1 and "Fig. 2".

Table 1: No. of trainees in each category

Class	X (BMI Ave.)	No. Of Trainees	Area (Quantity/n)	Y (Area/5)
1	17.5	59	0.0643	0.0129
2	22.5	213	0.2323	0.0465
3	27.5	359	0.3915	0.0783
4	32.5	201	0.2191	0.0583
5	37.5	60	0.0654	0.0131
6	42.5	14	0.0153	0.0031
7	47.5	7	0.0076	0.0015
8	52.5	4	0.0044	0.0009

©2020, IJA-ERA - All Rights Reserved

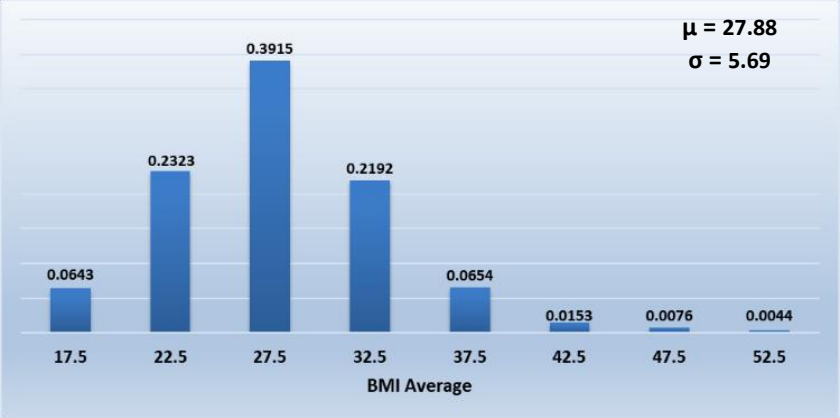

Figure 2: Trainees distribution depends on BMI

We also observe from "Fig. 2" that the distribution varied and therefore the mean and the standard deviation when we used BMI, where the mean and the standard deviation were 27.88 and 5.69 respectively. This certainly indicates the effect of selecting the appropriate standard to obtain the correct distribution of the selected sample.

B. Normal Distribution

As mentioned earlier, common mistakes in data distribution are the assumption that they are consistent with known statistical distributions. To prove the extent of the error, we assumed that the selected sample follows the normal distribution as assumed by the Military Training Institute (MTI) and we plot it as in "Fig. 3", using the normal distribution function by knowing mean and standard deviation as in (2)[5]. The black line represents the normal distribution and the vertical columns represent the actual distribution what we computed in II.A.

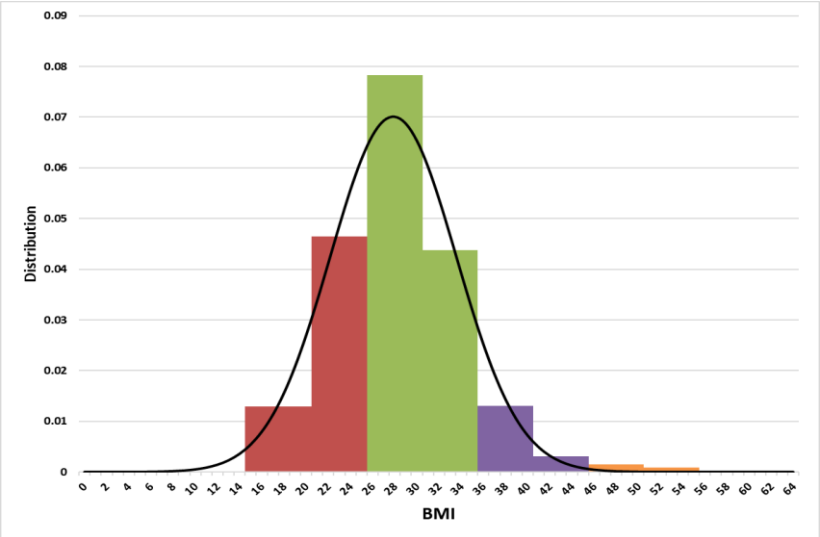

Figure 3: Normal & Actual Distribution of BMI

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

©2020, IJA-ERA - All Rights Reserved

C. *Regression line*

A regression line is simply a single line that best fits the data (in terms of having the minimum distance from the line to the points). The formula for the best-fitting line (or regression line) as in (3), where m is the slope of the line and b is the y-intercept [6].

$$y = mx + b \quad (3)$$

For our data, we have four regression lines that differ according to the BMI average distribution for each category as shown in "Fig. 4", which x-axis represent the BMI average and y-axis represent the area divided by 5 as shown in Table 1.

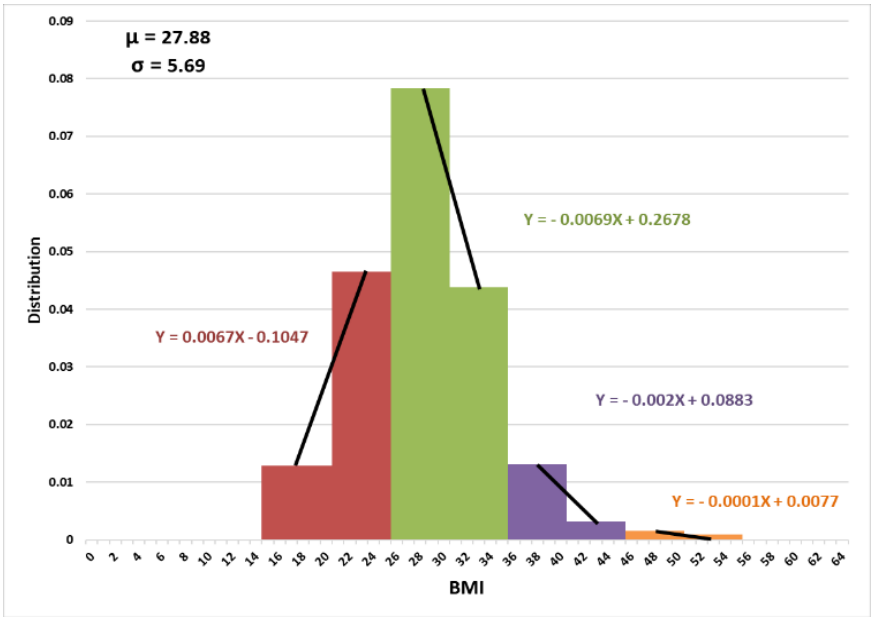

Figure 4: The Regression lines for each category

First and Second categories have an increasing regression line as in (4). Third and fourth categories have a decreasing regression line as in (5). Fifth and sixth categories have also decreasing regression line as in (6). Last two categories have a decreasing regression line as in (7).

$$y = 0.00672x - 0.1047 \quad (4)$$

$$y = -0.00690x + 0.2678 \quad (5)$$

$$y = -0.00200x + 0.0883 \quad (6)$$

$$y = -0.00012x + 0.0077 \quad (7)$$

D. *Applying the suitable formula*

The suitable formula will be used to identify the best fit for the original data, where each data has its formula. Our data has four different suitable formulas, where we have four different regression lines cover four different areas. First, we should compute the quadratic formula for our regression lines by calculating the integration of the equations in (4), (5), (6) and (7) [7-8]. For the equation in (4) the integration will be as in (8):

©2020, IJA-ERA - All Rights Reserved

$$F(x) = \int_{15}^{x} 0.00672t - 0.1047\, dt$$

$$= [\,0.00336t^2 - 0.1047]_{15}^{x}$$

$$F(x) = 0.00336x^2 - 0.1047x + 0.8145 \quad (8)$$

Now, the quadratic formula for the first regression line will be as in (9), which R is the random number between zero and 0.2966 (the total area of the first two categories from Table 1).

$$X = \frac{-(-0.1047) \pm \sqrt{(-0.1047)^2 - 4(0.00336)(0.8145 - R)}}{2\,(0.00336)} \quad (9)$$

For equations in (5), (6) and (7) the integrations formula will be as in (10), (11) and (12) respectively, which we should add the total areas of all the categories before it.

$$F(x) = -0.0034x^2 + 0.2680x - 4.2484 \quad (10)$$

$$F(x) = -0.0010x^2 + 0.0881x - 0.9512 \quad (11)$$

$$F(x) = -0.0001x^2 + 0.0072x + 0.7855 \quad (12)$$

Where the quadratic equations for (10), (11) and (12) will be as in (13), (14) and (15) respectively.

$$X = \frac{-(0.2680) \pm \sqrt{(0.2680)^2 - 4(-0.0034)(-4.2484 - R)}}{2\,(-0.0034)} \quad (13)$$

$$X = \frac{-(0.0881) \pm \sqrt{(0.0881)^2 - 4(-0.0010)(-0.9512 - R)}}{2\,(-0.0010)} \quad (14)$$

$$X = \frac{-(0.0072) \pm \sqrt{(0.0072)^2 - 4(-0.0001)(0.7855 - R)}}{2\,(-0.0001)} \quad (15)$$

Where R is a random number between 0.2966 & 0.9073 for the formula in (13) (0.9073 is the total area of the first fourth categories from Table 1). For the formula in (14), R is a random number between 0.9073 & 0.9879 (where 0.9879 is the total area of the first sixth categories from Table 1). Finally, R in the formula in (15) is between 0.9879 & 1 which 1 is the total area of all the categories in Table 1.

E. Generating new random numbers R

In this step, we will use the formulas we previously calculated for our original data to generate matching data by generating random numbers between 0 and 1 to compare how well they match the original data. As shown in "Fig. 5", we generated one million random numbers and applied them to our formulas using artificial intelligence. We found that they practically match the original data. As more random numbers you generate as you can get more match data of the original one.

 ©2020, IJA-ERA - All Rights Reserved

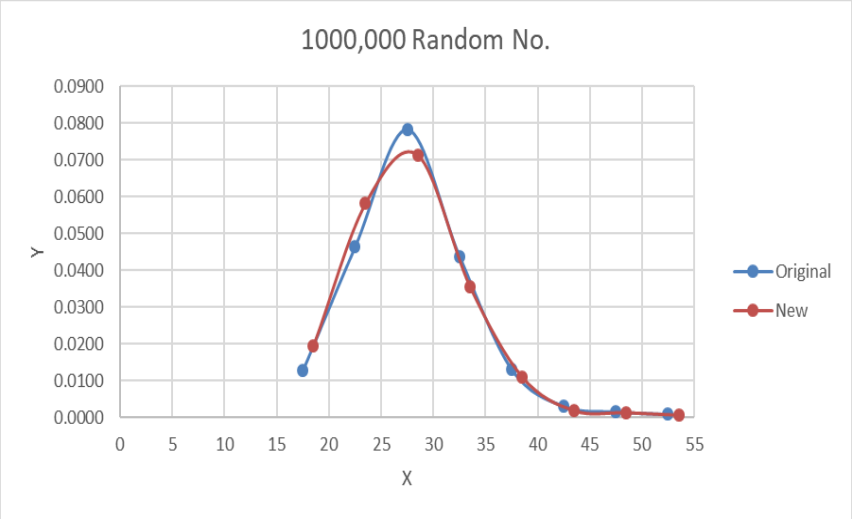

Figure 5: comparing original and new data

F. Comparison between assumed, fitted & actual distribution

As mentioned earlier, decisions based on the wrong assumption of data distribution will certainly lead to wrong results. So, we will show the superiority of our methodology over existing methods, measured as a percent error from the actual empirical data.

Table 2: Comparison between assumed & Actual distribution

X (BMI Ave.)	Actual Distribution %	Normalized Actual Distribution %	Assumed Distribution %	Normalized Assumed Distribution %	Difference Between Actual & Assumed %	Absolut (Difference Between Actual & Assumed) %
17.5	1.29	6.45	7.19	7.28	-0.83	0.83
22.5	4.65	23.24	22.47	22.74	0.50	0.50
27.5	7.83	39.13	33.58	33.98	5.15	5.15
32.5	4.38	21.89	25.01	25.31	-3.42	3.42
37.5	1.31	6.55	8.90	9.01	-2.46	2.46
42.5	0.31	1.55	1.53	1.55	0.00	0.00
47.5	0.15	0.75	0.12	0.12	0.63	0.63
52.5	0.09	0.45	0.01	0.01	0.44	0.44
	20.01	100	98.81	100		13.425

As shown to us in Table 2, the total difference between actual distribution and assumed distribution was 13.425 %. This ratio shows the amount of error and the distance between the assumed distribution and the actual distribution of data.

The total difference between actual distribution and fitted (polygonal) distribution was 0.00 % as shown in Table 3 which that mean they are identical. “Fig. 6” shows us the three distributions actual, assumed and fitted in one graph. Vertical bars represent the actual distribution, blue curve represent the assumed distribution (in our case is normal distribution) and the black lines represent the (polygonal) fitted distribution.

©2020, IJA-ERA - All Rights Reserved

I. ARTIFICIAL INTELLIGENCE

Artificial intelligence has helped us to facilitate the implementation of the previous steps with the push of a button and certainly to get very accurate results. We have built a program that only needs you to provide the data to be used and it will do all the calculations and extract the appropriate equations for them. In addition, the program will enable you to extract up to 1 million amount of data that exactly matching the distribution of your basic data as in "Fig. 7" and "Fig. 8".

Table 3: Comparison between fitted & Actual distribution

X (BMI Ave.)	Actual Distribution %	Normalized Actual Distribution %	Fitted Polygonal Distribution %	Normalized Assumed Distribution %	Difference Between Actual & Fitted Polygonal %	Absolut (Difference Between Actual & Fitted Polygonal) %
17.5	1.29	6.45	1.29	6.45	0.00	0.00
22.5	4.65	23.24	4.65	23.24	0.00	0.00
27.5	7.83	39.13	7.83	39.13	0.00	0.00
32.5	4.38	21.89	4.38	21.89	0.00	0.00
37.5	1.31	6.55	1.31	6.55	0.00	0.00
42.5	0.31	1.55	0.31	1.55	0.00	0.00
47.5	0.15	0.75	0.15	0.75	0.00	0.00
52.5	0.09	0.45	0.09	0.45	0.00	0.00
	20.01	100	20.01	100		0.00

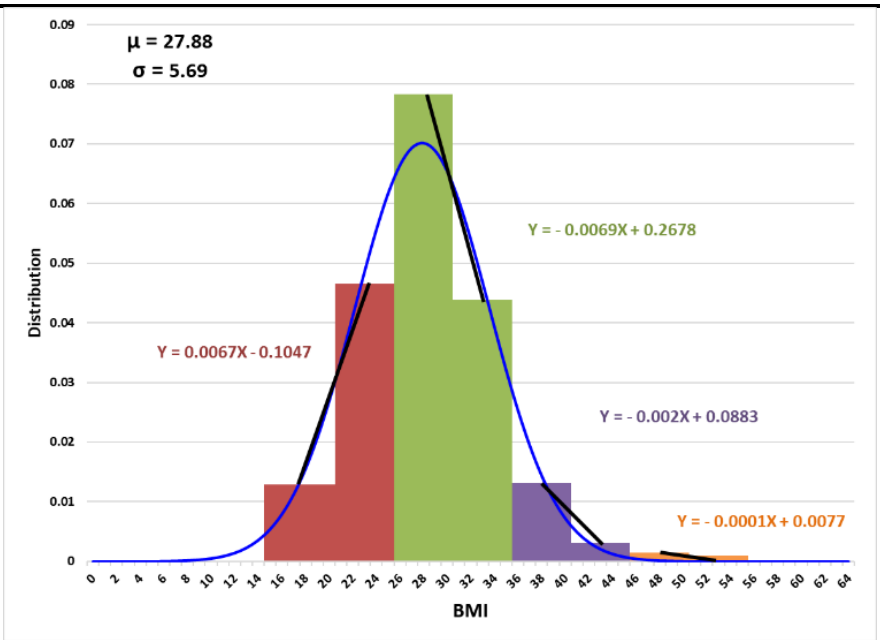

Figure 6: actual, assumed and fitted distribution

©2020, IJA-ERA - All Rights Reserved

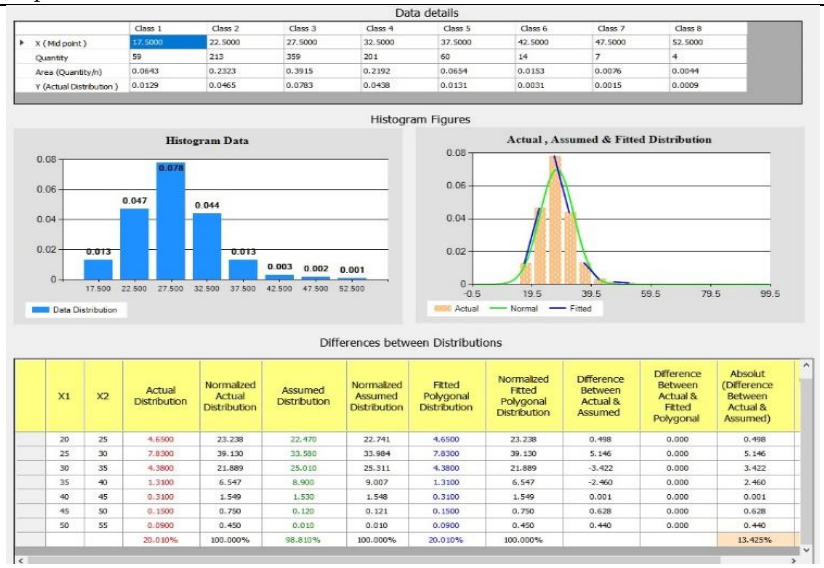

Figure 7: Data details, figures and differences table

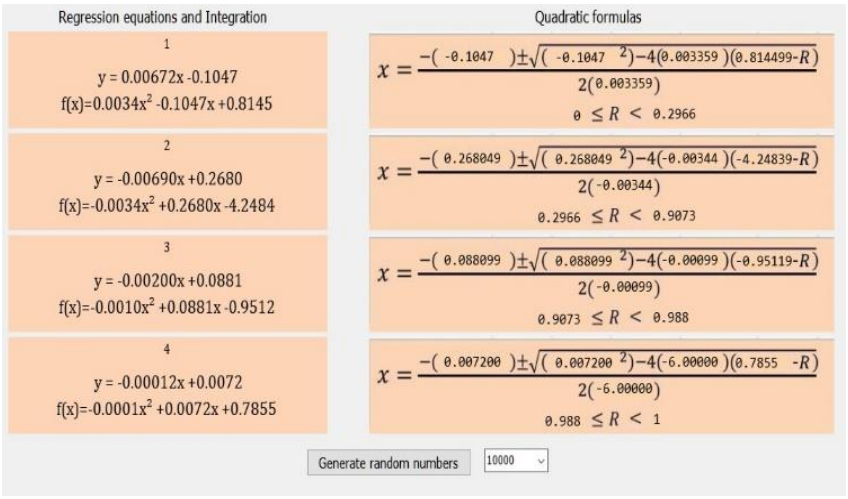

Figure 8: regression equations, quadratic formulas and generating random numbers

II. CONCLUSION

After detailed study and many experiments on different data sets using the new methodology that develops goodness of fit. We concluded that the new methodology has promising results with 0% error rate. Since the previous method used does not change no matter how the data used changes. We used artificial intelligence to analyze and select the lowest error rate and obtain the lowest and most accurate final formulas. Depending on the nature of the data, some formulas are expected to be immune to generalization across the range of the data.

REFERENCES

[1] Bury, K. (1999). Statistical distributions in engineering. Cambridge University Press.

[2] Ramberg, J. S., Dudewicz, E. J., Tadikamalla, P. R., & Mykytka, E. F. (1979). A probability distribution and its uses in fitting data. Technometrics, 21(2), 201-214.

 ©2020, IJA-ERA - All Rights Reserved

[3] MTI. Age, Height, Weight and Fitness rate records of 917 trainees. Military Training Institute. Jeddah, Saudi Arabia, 2018.

[4] Keys, A., Fidanza, F., Karvonen, M. J., Kimura, N., & Taylor, H. L. (1972). Indices of relative weight and obesity. Journal of chronic diseases, 25(6-7), 329-343.

[5] Ahsanullah, M., Kibria, B. G., & Shakil, M. (2014). Normal Distribution. In Normal and Student´ st Distributions and Their Applications (pp. 7-50). Atlantis Press, Paris.

[6] Yan, X., & Su, X. (2009). Linear regression analysis: theory and computing. World Scientific.

[7] Rassias, T. M. (1994). Topics in polynomials: extremal problems, inequalities, zeros. World Scientific.

[8] Blinn, H. (2005). How to solve a quadratic equation?. IEEE computer Graphics and Applications, 25(6), 76-79.

 ©2020, IJA-ERA - All Rights Reserved