HL Mathematics Internal Assessment

Finding the average shortest distance between any point in a unit square and the side

Candidate number: gbc594

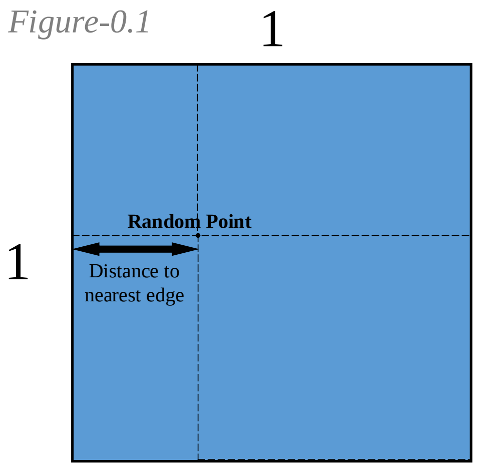
Table of Contents

Introduction

Background

When my family first moved to Singapore, my friends and I used to go swimming a lot in a large outdoor swimming pool that is approximately square-shaped. There is a lightning alarm beside the pool that would be activated whenever it detects a thunder storm approaching. The life guard told us that everyone has to exit the pool whenever the lightning alarm is triggered, for that the pool containing salt water is very likely to be struck by lightning. This cause a lot of panic, as my friends and I would race to the nearest side of the pool whenever the lightning alarm is triggered to get out of the pool as quickly as possible. Although I've never seen or heard the pool being struck by lightning in the past 6 years, I have always wondered how much distance on average my friends and I would have to cover to exit the pool from the closest side before a lightning strike.

Figure-0.1



Refining the Problem

To simplify this problem, the position of my friends and I in the pool is assumed to be randomly distributed, and the pool is approximated to a unit square. This refines the aim of this exploration to finding the average shortest distance between any point in a unit square and the closest side, illustrated by *Figure-0.1*. There is a theoretical and an empirical approach to fulfill the aim: The theoretical approach involves calculating an exact solution to the problem through calculus, and the empirical approach involves randomly generating a large number of coordinates within the unit square and averaging their shortest distance to the closest side through programming. This exploration will attempt to fulfill its aim from both approaches and compare the extent to which they agree with each other. This comparison will be meaningful as it can reveal the fundamental nature of calculus and statistics.

Aim (Restated)

To find the average shortest distance between any point in a unit square and the side through two approaches and compare their results.

Part I: The Theoretical (Calculus) Approach

Step 1: Geometric analysis of the problem

The first thing to note about this problem is that the shortest distance to any side in a square will always be the perpendicular distance to the side. This can be easily proved by Pythagoras's theorem, as any distance between a point and a side that is not perpendicular to the side (c) can always be regarded as the hypotenuse of a right triangle made from the perpendicular distance (b) and the side (a), as seen on *Figure-1.1*. Since that the Pythagoras's theorem states that:

$$c^2 = a^2 + b^2$$

, a non-perpendicular distance will always be longer than a perpendicular one between a point and a side.

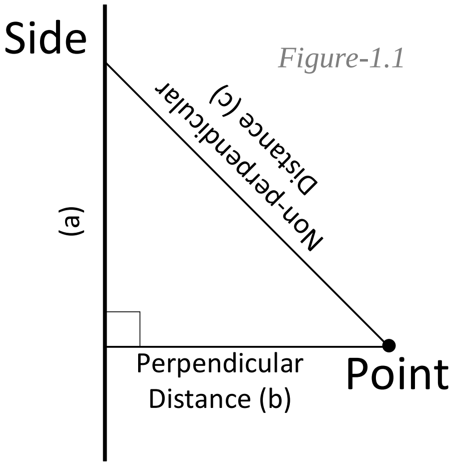


Figure-1.1

Given that a square has four sides in total, any point in a square will have four different perpendicular distances between it and each of the four sides, but only one of them will be the shortest distance. Instead of considering all four sides at once, we can focus on one side at a time if we split the unit square into four analogous quarter triangles as shown on *Figure-1.2A*. As on *Figure-1.2A*, if a point falls within the area of Quarter Triangle A, it can be said for definite that this point will have the shortest perpendicular distance with Side A, and the same can be said about Side-Triangle pairs B, C and D.

Figure-1.2A

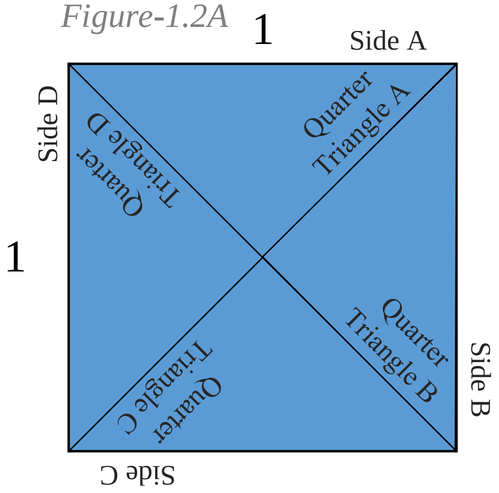


Figure-1.2B

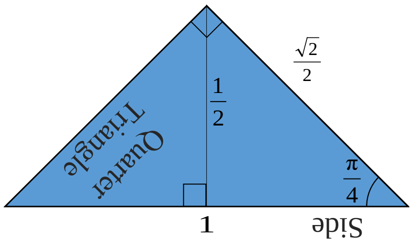


Figure-1.2B shows the dimensions of a single quarter triangle. Since that each of the four quarter triangles are completely analogous, the probability that a random point in the unit square falling within the bounds of any of the four quarter triangles are equally likely. This means that the average distance between any point within a quarter triangle and the side it shares with unit square is equivalent to the average shortest distance between any point in the unit square and the closest side, hence the problem can be simplified by considering a single quarter triangle instead of the whole unit square.

A quarter triangle can be modeled as follows on a Cartesian plane, as seen on *Figure-1.3*:

$$y \leq -x + \frac{\sqrt{2}}{2}$$

$$x \in \left[0, \frac{\sqrt{2}}{2}\right]$$

$$y \in \left[0, \frac{\sqrt{2}}{2}\right]$$

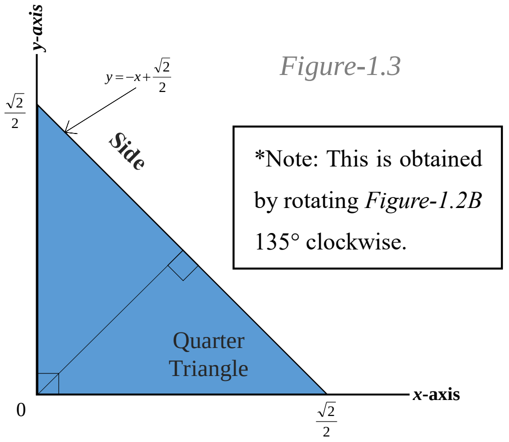


Figure-1.3

*Note: This is obtained by rotating *Figure-1.2B* 135° clockwise.

Step 2: Finding the probability of randomly selecting a specific point in the triangle

To find the exact average of the distances between any point and the side, all possible points within the quarter triangle will have to be considered. One way to do this is to find an expression for the probability of randomly selecting a specific point out of all possible points in the quarter triangle.

To find the probability of randomly selecting a specific point in the quarter triangle, we should first consider the probability of selecting a random point in the square formed by the domain and range of $y \in \left[0, \frac{\sqrt{2}}{2}\right]$ and $x \in \left[0, \frac{\sqrt{2}}{2}\right]$. Let Δx and Δy be any continuous length of x and y values within the restricted domain and range, as represented *Figure-1.4A*. The probability of randomly selecting a point that falls within the length of Δx or Δy are respectively $\frac{\Delta x}{\sqrt{2}/2}$ and $\frac{\Delta y}{\sqrt{2}/2}$, hence the probability of randomly selecting a point that falls within the rectangular area formed by Δx and Δy, "P(Area)", can be expressed as:

$$\text{P(Area)} = \frac{\Delta y}{\sqrt{2}/2} \times \frac{\Delta x}{\sqrt{2}/2}$$

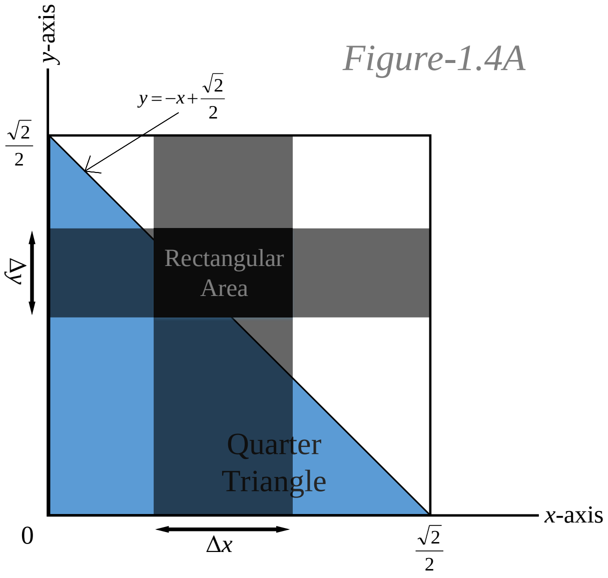


Figure-1.4A

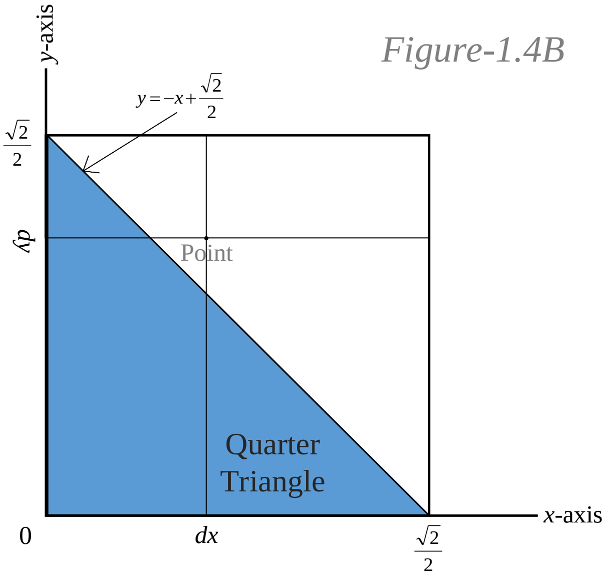


Figure-1.4B

Figure-1.4B shows that when distances Δx and Δy approach infinitesimal and become differentials dx and dy, the rectangular area they form will approach a specific point. Therefore, the probability of randomly selecting a single point, "P(Point)", in the restricted domain and range, would be:

$$\mathrm{P(Point)} = \frac{dy}{\sqrt{2}/2} \times \frac{dx}{\sqrt{2}/2}$$

Since the area of the quarter triangle is half of the restricted domain and range, there is half as many possible positions that can be randomly selected in the quarter triangle than the restricted domain and range, which also means that the probability of randomly selecting a specific point in the quarter triangle will be twice of the probability of random selecting a specific point within the domain and range. P(Point) in a quarter triangle is hence calculated as:

$$\boxed{\mathrm{P(Point)} = 2\left(\frac{dy}{\sqrt{2}/2} \times \frac{dx}{\sqrt{2}/2}\right)}$$

Step 3: Finding the perpendicular distance between any random point and the side

After finding the probability of each point being selected in a quarter triangle, we need to find an expression for the distance between any point in the quarter triangle and the side of the unit square.

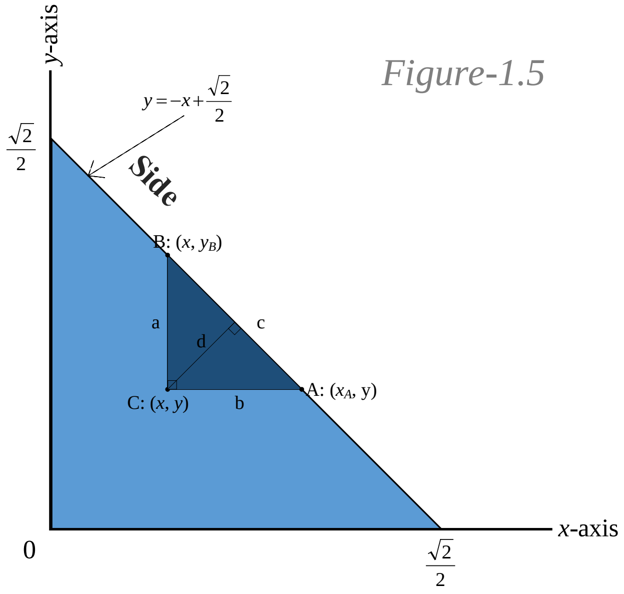


Figure-1.5

Let a random point C on *Figure-1.5* have the coordinates (x, y). Distance "d" is the perpendicular distance between point C and the side, the goal here is to find an expression for distance d in terms of x and y.

Suppose point C have the horizontal distance "b" and the vertical distance "a" from the side, extending to point B and point A to make triangle ABC. With information derived from *Figure-1.5*, the lengths of triangle ABC's three sides "a", "b", and "c" can be expressed in terms of x and y as follows:

$$\mathrm{a} = |y - y_B|$$

$$\mathrm{a} = \left|y - \left(-x + \tfrac{\sqrt{2}}{2}\right)\right|$$

$$\mathrm{a} = \left|x + y - \tfrac{\sqrt{2}}{2}\right|$$

$$\mathrm{b} = |x - x_A|$$

$$\mathrm{b} = \left|x - \left(-y + \tfrac{\sqrt{2}}{2}\right)\right|$$

$$\mathrm{b} = \left|x + y - \tfrac{\sqrt{2}}{2}\right|$$

$$\boxed{\mathrm{a} = \mathrm{b} = \left|x + y - \frac{\sqrt{2}}{2}\right|}$$

$$\mathrm{c} = \sqrt{\mathrm{a}^2 + \mathrm{b}^2}$$

$$\mathrm{c} = \sqrt{2\mathrm{a}^2}$$

$$\boxed{\mathrm{c} = \sqrt{2}\left|x + y - \frac{\sqrt{2}}{2}\right|}$$

There are two ways to calculate the area of triangle *ABC*:

$$\boxed{\begin{gathered}\text{Area} = \frac{1}{2}(\mathrm{c} \times \mathrm{d}) \\ \text{Area} = \frac{1}{2}(\mathrm{a} \times \mathrm{b})\end{gathered}}$$

We can hence equate the two ways of finding the same area in terms of *x* and *y* to calculate distance "d" as follows:

$$\frac{1}{2}(c \times d) = \frac{1}{2}(a \times b)$$

$$\sqrt{2}\left|x + y - \frac{\sqrt{2}}{2}\right| \times d = \left(x + y - \frac{\sqrt{2}}{2}\right)^2$$

$$d = \frac{\left|x + y - \frac{\sqrt{2}}{2}\right|}{\sqrt{2}}$$

Given that the area is restricted to $\mathrm{y} \leq \frac{\sqrt{2}}{2} - x$:

$$\boxed{d = \frac{\frac{\sqrt{2}}{2} - (x + y)}{\sqrt{2}}}$$

Step 4: Finding the average perpendicular distance between all possible points and the side in a quarter triangle

To find the mean or expected value of any random variable, we take the sum of the products between all possible values of the random variable and their respective probability of occurring. From previous steps, we know how to find the probability of any specific point being selected and how to find distance *d* for any specific point; Hence the product between any possible value of the continuous random variable *d* and its probability of occurring can be expressed as:

$$\left[\frac{\frac{\sqrt{2}}{2} - (x + y)}{\sqrt{2}}\right]\left(2 \times \frac{dy}{\sqrt{2}/2} \times \frac{dx}{\sqrt{2}/2}\right)$$

To find the expected value of *d*, “E(*d*)”, we put the product expression into a double integral with respect to both *x* and *y* to calculate the sum of the product of all possible *d* values and their probability of occurrence for all possible pairs of *x* and *y* values within the area of the quarter triangle. The first integral with respect to *x* will have the same bounds as the domain of the quarter triangle model to make sure that only *x* values that fall into the quarter triangle are considered. The bounds of the second integral with respect to *y* will change with respect to *x* given the inequality $y \leq -x + \frac{\sqrt{2}}{2}$ that models the quarter triangle, for that the possible *y* values changes at different *x* values. This finally gives the expression:

$$\boxed{\mathrm{E}(d) = \int_0^{\frac{\sqrt{2}}{2}} \int_0^{\frac{\sqrt{2}}{2}-x} \left[\frac{\frac{\sqrt{2}}{2} - (x+y)}{\sqrt{2}}\right] \left(2 \times \frac{dy}{\sqrt{2}/2} \times \frac{dx}{\sqrt{2}/2}\right)}$$

This expression can be simplified and solved as follows:

$$\mathrm{E}(d) = \int_0^{\frac{\sqrt{2}}{2}} \int_0^{\frac{\sqrt{2}}{2}-x} \left[\frac{\sqrt{2}}{2} - (x+y)\right] \; 2\sqrt{2} \; dy \; dx$$

$$\mathrm{E}(d) = 2\int_0^{\frac{\sqrt{2}}{2}} \int_0^{\frac{\sqrt{2}}{2}-x} \left[(1-\sqrt{2}x) - \sqrt{2}y\right] \; dy \; dx$$

$$\mathrm{E}(d) = 2\int_0^{\frac{\sqrt{2}}{2}} \left[(1-\sqrt{2}x)y - \frac{\sqrt{2}}{2}y^2\right]_0^{\frac{\sqrt{2}}{2}-x} \; dx$$

$$\mathrm{E}(d) = 2\int_0^{\frac{\sqrt{2}}{2}} \left[(1-\sqrt{2}x)\left(\frac{\sqrt{2}}{2} - x\right) - \frac{\sqrt{2}}{2}\left(\frac{\sqrt{2}}{2} - x\right)^2\right] \; dx$$

$$\mathrm{E}(d) = 2\int_0^{\frac{\sqrt{2}}{2}} \left[\sqrt{2}\left(\frac{\sqrt{2}}{2} - x\right)^2 - \frac{\sqrt{2}}{2}\left(\frac{\sqrt{2}}{2} - x\right)^2\right] \; dx$$

$$\mathrm{E}(d) = \sqrt{2}\int_0^{\frac{\sqrt{2}}{2}} \left(x^2 - \sqrt{2}x + \frac{1}{2}\right) \; dx$$

$$\mathrm{E}(d) = \sqrt{2}\left[\frac{x^3}{3} - \frac{\sqrt{2}}{2}x^2 + \frac{1}{2}x\right]_0^{\frac{\sqrt{2}}{2}}$$

$$\mathrm{E}(d) = \sqrt{2}\left[\frac{1}{3}\left(\frac{\sqrt{2}}{2}\right)^3 - \frac{\sqrt{2}}{2}\left(\frac{\sqrt{2}}{2}\right)^2 + \frac{1}{2}\left(\frac{\sqrt{2}}{2}\right)\right]$$

$$\boxed{\mathrm{E}(d) = \frac{1}{6}}$$

From the result of solving this equation, it can be hence concluded for Part I that the average shortest distance between any point in a unit square and the side is $1/6$ units.

Reflection on Part I

There is a minor "error", or rather, imperfection in the method used in this part of the investigation. With reference to *Figure-1.6A* and *Figure-1.6B*, splitting the unit square into four quarter triangles and considering them independently will cause all the points along the cleavages of quarter triangles to be double-counted in the final calculation, as these points are in fact simultaneously closest to two or more sides instead of only one as assumed in Step 1.

Figure-1.6A

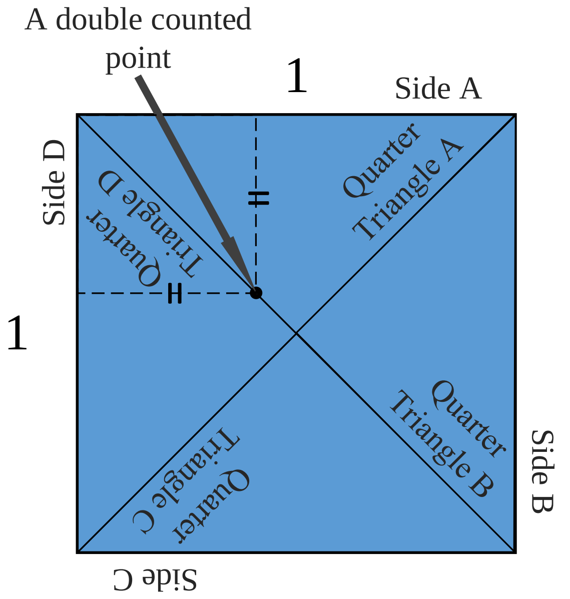
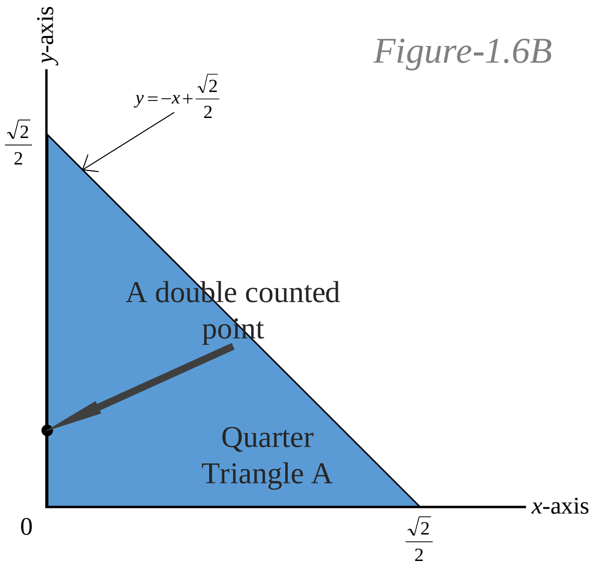


Figure-1.6B



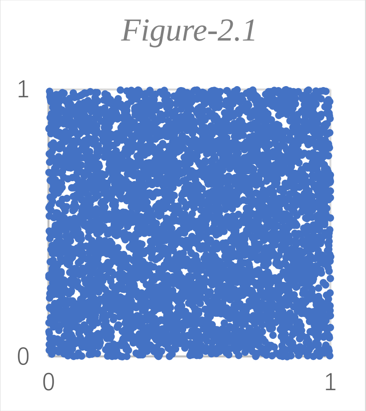
However, one should note that the method used in Part I remains valid in terms of achieving the aim, as the area of the cleavage is infinitely small compared to the area of the unit square, which hence makes the relative error caused by double counting infinitely small. On the other hand, having an infinitely small error does not mean there is no error, hence the result obtained from Part I cannot be taken as "exactly accurate".

In the end, with the purpose of this investigation being to solve a real-life problem, it is only meaningful to omit the infinitely small error and assume that $1/6$ is the exact solution to the problem, for that we cannot take an infinitely small measurement of the length of a swimming pool in practice.

Part II: The Empirical (Statistics) Approach

Methodology and result

5000 random coordinates are generated in a unit square through Microsoft excel, shown on *Figure-2.1*. The perpendicular distance between each point and all four sides are then calculated, from which the minimum of the four distances is taken as the minimum distance to the closest side.



In terms of notation, let the letter "*n*" represent the n^{th} random position generated in sequence out of the 5000 positions, where each random position's minimum distance to the closest side is represented by "d_n" with "*n*" being its order of occurrence, and let "μ_n" represent average shortest distance to the closest side for the sample up to the n^{th} random position generated.

"μ_n" is calculated by the following equation:

$$\boxed{\mu_n = \frac{\sum_{i=1}^{n} d_i}{n}}$$

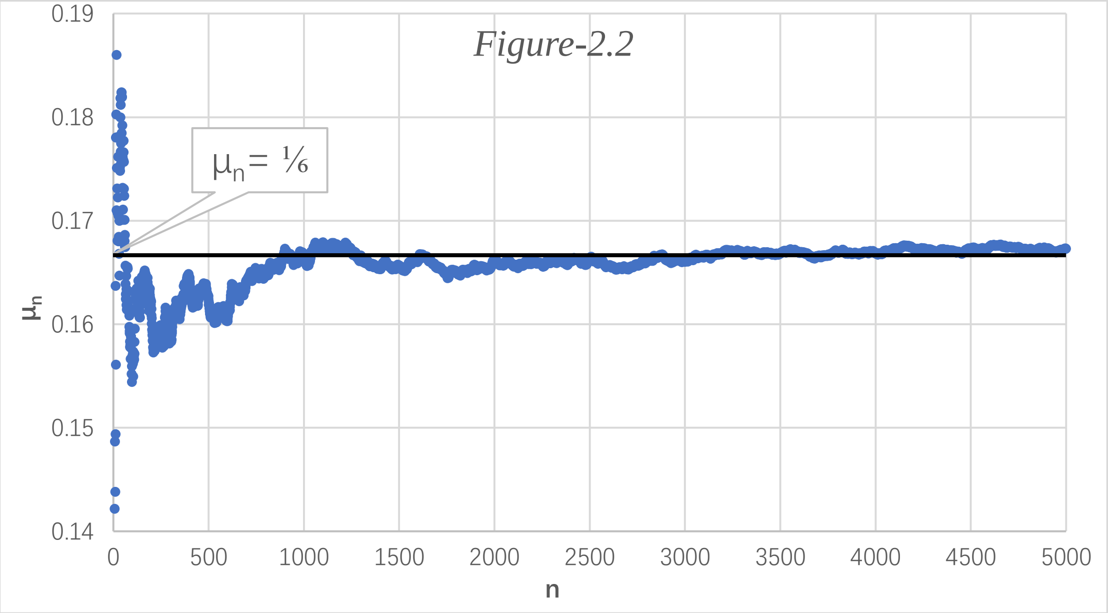
After 5000 random positions have been generated, the following result is obtained:

$$\boxed{\mu_{5000} \approx 0.1673}$$

The value 0.1673 deviates from the exact solution of $1/6$ calculated in Part I by approximately 0.4%. Although the discrepancy is relatively small, it can't yet be concluded that the results from part I and part II agree with each other. To confirm that the result of part I and part II are concordant, the behavior of μ_n as n increases will have to be analyzed.

Analysis of behavior and conclusion

The behavior of μ_n as n increases up to 5000 is illustrated on *Figure-2.2*.



It can be observed on *Figure-2.2* that the value of μ_n constantly fluctuates as the value of n increases. While the magnitude of fluctuation started off to be relatively large, it gradually decreases as the value of n increases, where μ_n eventually converges towards $1/6$. From this observed behavior we can make the conjecture that:

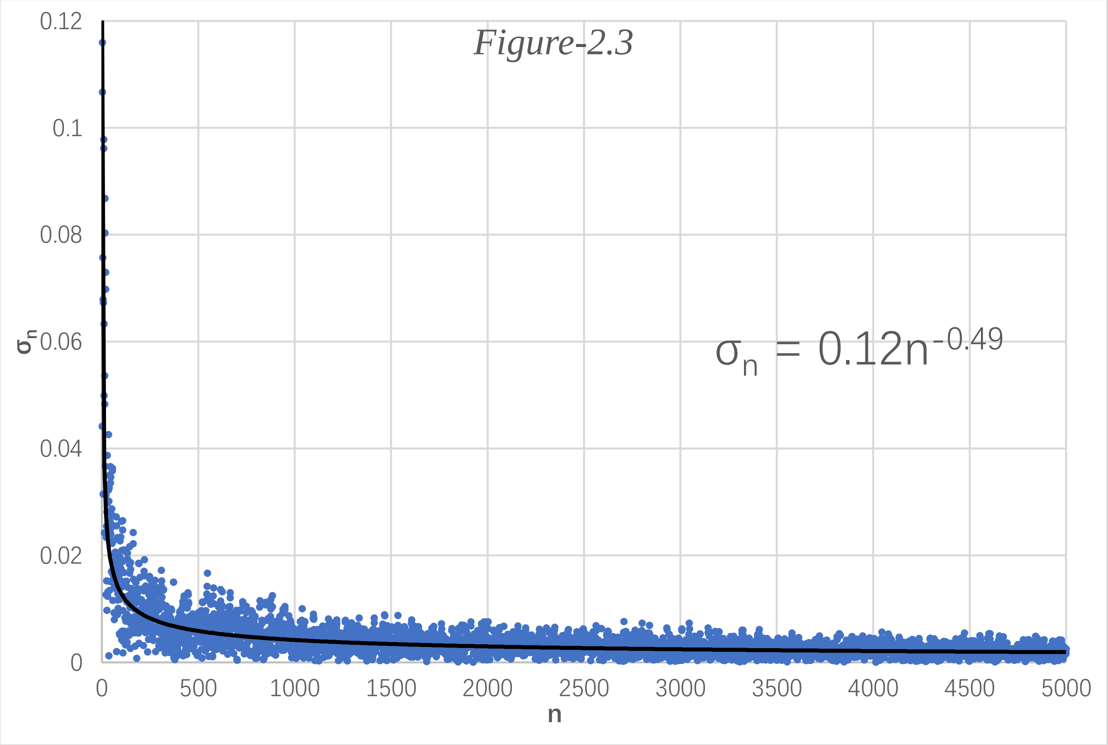
$$\lim_{n\to\infty} \mu_n = \frac{1}{6}$$

To justify this conjecture, we can look at how the standard deviation between the sample and the value $1/6$ behaves as the sample size of random positions increases. Let "σ_n" represent the standard deviation between $1/6$ and the sample from d_1 to d_n; σ_n is calculated by the following formula:

$$\sigma_n = \sqrt{\frac{\sum_{i=1}^{n} \left(d_i - \frac{1}{6}\right)^2}{n}}$$

Note that although the sample standard deviation formula would seem more appropriate in this situation, where a finite sample of $n \leq 5000$ is taken from infinite possible samples, I have decided to calculate "σ_n" as a population standard deviation. This is because the purpose of calculating the standard deviation here is to evaluate the how much a sample deviates from the specific value of $1/6$ instead of its mean. The sample standard deviation is adjusted for the fact that the mean changes as the sample changes, but the value $1/6$ remains fixed no matter how the sample changes, hence that the population standard deviation is more appropriate here.

Figure-2.3 illustrates the behavior of σ_n as the sample size n increases.



It is evident from *Figure-2.3* that the larger the sample size of random positions in a unit square, the less the average distance to the closest side will deviate from the value $1/6$. *Figure-2.3* also shows that the relationship between σ_n and n can be modeled by the equation $\sigma_n = 0.12n^{-0.49}$ within the given domain. This equation has a horizontal asymptote at $\sigma_n = 0$, which means that as the sample size gets infinitely large, the standard deviation between the sample distances and the value $1/6$ will become infinitely small, hence:

$$\boxed{\lim_{n\to\infty} \sigma_n = 0}$$

Since the standard deviation is a measure of spread, if the standard deviation between the sample and $1/6$ becomes infinitely small, the mean of the sample will inevitably converge towards the value $1/6$, which strongly justifies the previous conjecture that:

$$\boxed{\lim_{n\to\infty} \mu_n = \frac{1}{6}}$$

From this justified conjecture, we can hence conclude for part II that the statistics approach has given sufficient evidence that the average shortest distance between any point in a unit square and the closest side is $1/6$ units. This conclusion strongly agrees with the conclusion of Part I.

Reflection on Part II

The conclusion drawn through the empirical approach will only remain as a conjecture instead of being an actual solution, as the claim that $\lim_{n\to\infty} \mu_n = \frac{1}{6}$ could not be definitively proven or verified from the data generated. Firstly, this claim assumes that μ_n will continue to behave as observed from *Figure-2.2* beyond $n = 5000$, but a proven solution should not be based on assumptions. Secondly, it can't be proven for definite that μ_n converge towards exactly $1/6$ as n increases on *Figure-2.2*, while the equation $\sigma_n = 0.12n^{-0.49}$ does not exactly fit through all the data points on *Figure-2.3*; A somewhat similar behavior can still be observed from *Figure-2.2* and *Figure-2.3* if the value that μ_n converges toward is some other value that is close to $1/6$.

It is also very important to note that Part II of the investigation relies greatly on the conclusion of Part I. Without the exact solution calculated in Part I, it is impossible to even make the conjecture that $\lim_{n\to\infty} \mu_n = \frac{1}{6}$ in Part II.

Conclusion

Interpretation of results

The results of both approaches suggest that the average shortest distance between any point in a unit square and the side is $1/6$ units. The empirical result from Part II after 5000 random positions agrees with the exact solution from Part I with a relatively low discrepancy of approximately 0.4%. With the calculated exact solution from Part I being $1/6$, the conjecture based on observation that $\lim\limits_{n\to\infty} \mu_n = \frac{1}{6}$ in Part II confirms the postulate that the greater the sample size gets, the closer the sample mean gets to the theoretical expected value.

Taking these results back into the original problem, it can be said that the average distance that my friends and I would have to cover to exit the pool from the closest side would be one sixth the length of the side of the pool. Each side of the pool is 15 meters in length, hence on average we would be 2.5 meters away from the closest edge of the pool. My average swimming speed is around 1.4 meters per second (value calculated from my performance in a 50 meter race), which means I will take approximately 1.8 seconds on average to exit pool after the lightning alarm is triggered.

Implication of results

Since both approaches arrive at the same conclusion that the average is $1/6$ units, it can be established that:

$$\lim_{n\to\infty} \mu_n \cong \int_0^{\frac{\sqrt{2}}{2}} \int_0^{\frac{\sqrt{2}}{2}-x} \left[\frac{\frac{\sqrt{2}}{2} - (x+y)}{\sqrt{2}}\right] \left(2 \times \frac{dy}{\sqrt{2}/2} \times \frac{dx}{\sqrt{2}/2}\right)$$

This suggests that the calculus approach and $\lim\limits_{n\to\infty} \mu_n$ is somewhat similar in nature, which is true to a certain extent if we look at the mechanism of each of the approaches. As previously discussed, the only way to find the exact average of a continuous variable would be to consider all of the infinite possibilities. For the shortest distance between any point in a unit square and the closest side, all possible coordinates in the unit square will have to be considered when calculating the average. The calculus approach attempts to do so in an orderly fashion by considering all possible *x* values in sequence, while considering all the possible *y* values at each *x* values. On the other hand, the statistic approach also attempts to do so by generating random positions until all possible coordinates are accounted for. When all the possible coordinates are considered, the order at which they have been considered becomes unimportant, hence justifying that the conjecture of $\lim\limits_{n\to\infty} \mu_n$ and the calculus approach are similar in terms of what they try to achieve.

While both approaches attempt to take all possible coordinates in the unit square into account in the average calculation, the empirical approach inevitably fails to do so as it is impossible to randomly generate infinite numbers of coordinates within a finite period of time. However, the calculus approach succeeds in taking account of all coordinates by evaluating how the infinitely many points are distributed in terms of x and y instead of tallying each individual point.

Evaluation

Possible Improvements

The choice of generating only 5000 random positions in Part II is based on the processing capacity of my laptop instead of any mathematical reason. As the conclusion of Part II is a conjecture based on the data generated, generating more data through the empirical approach would make the conclusion of Part II both more meaningful and more reliable if the same behavior persists beyond $n = 5000$.

The sample average of the closest distance between 5000 random positions in the unit square and the nearest side is rounded to the fourth decimal place in Part II. The accuracy of four decimal places reveals two discrepant digits from the rounded exact solution (0.1673 and 0.1667) to calculate the percentage discrepancy effectively. However, the random number generator in Microsoft Excel keeps 15 decimal places of accuracy, which means that the result of Part II could have had 15 decimal places of justified accuracy.

As Part II of the investigation involves random number generator, the exact data set used for this investigation is impossible to replicate. To prove the observation that the average converging towards $1/6$ in Part II is not exclusive to this specific set of data, I could have generated multiple sets of 5000 random positions and graphically illustrate how they all converge towards $1/6$ like I did on *Figure-2.2*. This would further justify the conjecture that $\lim_{n\to\infty} \mu_n = 1/6$, and make the conclusion of this investigation more reliable.

Extension to the investigation

The ideas explored in this investigation lends itself to further explorations that involves finding the relationship between a dependent and an independent variable. For example, a meaningful extension of this investigation would be to investigate how the average shortest distance from the closest side in a rectangular swimming pool would change as its dimensions change, as most professional swimming tracks are in fact rectangular in shape. Focusing more on the geometric significance of this investigation, another interesting extension would be to investigate the relationship between the number of sides that a regular polygon has and the average shortest distance between any point in the polygon and the closest side.