The Average Distance between Two Points

Preliminaries:

1) The expected value $\mathbb{E}$.
2) Transformation of random variables
3) Double integrals
4) The Central Limit Theorem

Definition:

Denote $\Delta(\mathrm{n})$ as the expected Euclidean distance between two points uniformly and independently chosen in an n-dimensional unit hypercube.

Problem 1:

Find $\Delta(1)$. In this case, a 1-dimensional unit hypercube is a line segment of length 1.

Solution:

Suppose the left end of the line segment is at the origin of the real number line, and the right end is at 1.

Consider two independent standard uniform random variables U and V that equal the values of the two points on the real number line. Let D be the distance between U and V.

$$\mathrm{D} = |\mathrm{U} - \mathrm{V}|$$

Our goal is to find (D) .

The probability density function of D is:

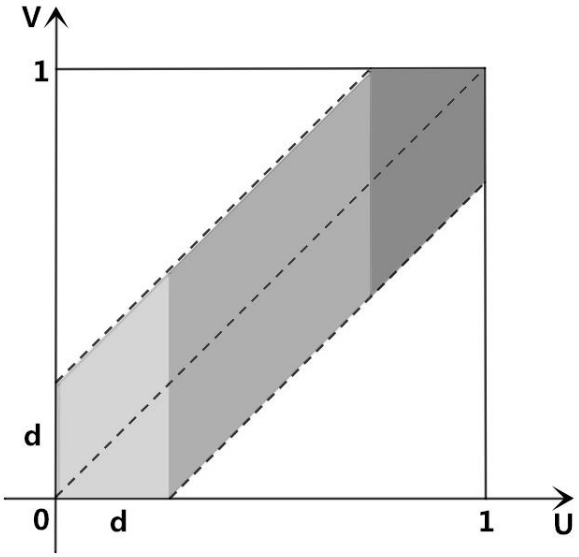
$$\mathrm{F_D(d)} = \mathrm{P(D \le d)} = \mathrm{P}(|\mathrm{U} - \mathrm{V}| \le \mathrm{d})$$

$$= \iint_A f(u,v)\, dvdu$$

(A is the region where $|\mathrm{U} - \mathrm{V}| \le \mathrm{d},\ 0 \le \mathrm{U}, \mathrm{V} \le 1$)

$$= \iint_A f(u)f(v)\, dvdu$$ (since U and V are independent)

$$= \iint_A 1\, dvdu$$



We are integrating the shaded region. We will divide the entire region into three smaller regions, integrate each separately, and add them together.

$$\mathrm{F_D(d)} = \int_0^d \int_0^{u+d} 1 \, dvdu + \int_d^{1-d} \int_{u-d}^{u+d} 1 \, dvdu + \int_{1-d}^1 \int_{u-d}^1 1 \, dvdu$$

$$= \frac{3}{2}\mathrm{d}^2 + 2\mathrm{d} - 4\mathrm{d}^2 + \frac{3}{2}\mathrm{d}^2$$

$$= 2\mathrm{d} - \mathrm{d}^2 \qquad (0 < \mathrm{d} < 1)$$

$$F_{\mathrm{D}}(d) = \begin{cases} 0 & \mathrm{d} < 0 \\ 2\mathrm{d} - \mathrm{d}^2 & 0 \le \mathrm{d} \le 1 \\ 1 & \mathrm{d} > 1 \end{cases}$$

$$f_{\mathrm{D}}(d) = \frac{\mathrm{d}}{\mathrm{d}d} F_{\mathrm{D}}(d) = \begin{cases} 2 - 2\mathrm{d} & 0 \le \mathrm{d} \le 1 \\ 0 & otherwise \end{cases}$$

Thus,

$$\mathbb{E}(\mathrm{D}) = \int_{-\infty}^{\infty} d \cdot f_{\mathrm{D}}(d) dd = \int_0^1 \mathrm{d}(2 - 2\mathrm{d}) dd = \frac{1}{3}$$

Therefore, $\Delta(1) = \frac{1}{3} \approx 0.3333$.

□

Note that knowing the distribution of D will be useful to solving the next problem. The distribution of D is known as the triangular distribution. We could simplify the solution by directly solving for $\mathbb{E}(|U - V|)$ using $\mathrm{f(u, v)}$. This yields the alternate solution below.

Alternate Solution:

$$\mathbb{E}(|U - V|) = \iint_{all\ (u,v)} f(u, v)|u - v| \, dvdu$$

$$= \int_0^1 \int_0^1 |u - v| \, dv \, du$$

$= \int_0^1 \int_0^u (u - v)\, dv\, du + \int_0^1 \int_u^1 (v - u)\, dv\, du$

(the first integral for $u > v$, and the second integral for $u \leq v$)

$= \frac{1}{6} + \frac{1}{6}$

$= \frac{1}{3} \approx 0.3333$

□

Problem 2:

Find $\Delta(2)$.

Solution:

Consider two independent random variables D_1 and D_2. D_1 equals the absolute difference of the x-coordinates of the two points. D_2 equals the absolute difference of the y-coordinates of the two points. Then, both D_1 and D_2 have the same distribution as D .

Our goal is to find $(\sqrt{{D_1}^2 + {D_2}^2})$.

$\mathbb{E}\left(\sqrt{{D_1}^2 + {D_2}^2}\right) = \iint_{all\ (d_1, d_2)} f(d_1, d_2)\, \mathrm{d}d_1 \mathrm{d}d_2$

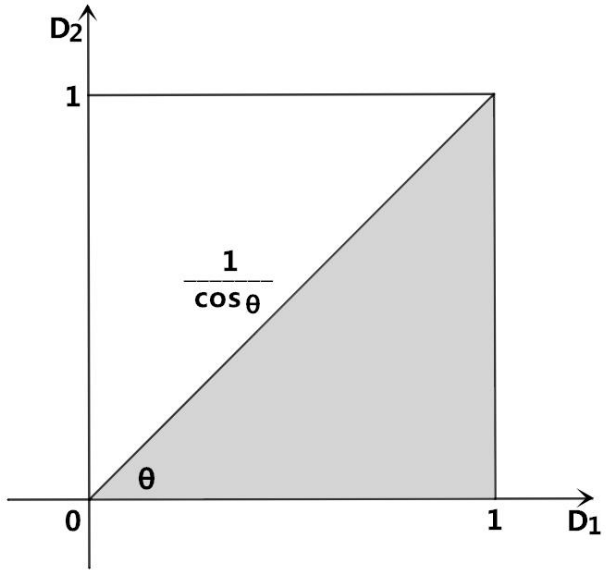
$= \iint_{all\ (d_1, d_2)} f(d_1) f(d_2)\, \mathrm{d}d_1 \mathrm{d}d_2$

(since D_1 and D_2 are independent)

$= \int_0^1 \int_0^1 (2 - 2d_1)(2 - 2d_2) \sqrt{x^2 + y^2}\, \mathrm{d}d_1 \mathrm{d}d_2$

$= 4 \int_0^1 \int_0^1 (1 - d_1)(1 - d_2) \sqrt{x^2 + y^2}\, \mathrm{d}d_1 \mathrm{d}d_2$

Now we changes variables to polar coordinates. Let $d_1 = \mathrm{r}\cos\theta$, $d_2 = \mathrm{r}\sin\theta$ $(\mathrm{r} \in \mathbb{R})$. The new integral has an additional factor of r.



Originally, we were integrating a square. Both d_1 and d_2 ranges from 0 to 1. Instead of integrating the entire square, we now integrate the shaded triangle and multiply it by 2. Then θ ranged from 0 to $\frac{\pi}{4}$, and r ranges from 0 to $\frac{1}{\cos\theta}$.

$$\mathbb{E}\left(\sqrt{{D_1}^2+{D_2}^2}\right)$$

$$= 4\int_0^{\frac{\pi}{4}} 2\int_0^{\frac{1}{\cos\theta}} (1-r\cos\theta)(1-r\sin\theta)\sqrt{(r\cos\theta)^2+(r\sin\theta)^2}\; r\, dr d\theta$$

$$= 8\int_0^{\frac{\pi}{4}} \int_0^{\frac{1}{\cos\theta}} (1-r\cos\theta)(1-r\sin\theta)\sqrt{r^2(\sin^2\theta+\cos^2\theta)}r\, dr\, d\theta$$

$$= 8\int_0^{\frac{\pi}{4}} \int_0^{\frac{1}{\cos\theta}} r^2(1-r\cos\theta)(1-r\sin\theta)\, dr\, d\theta$$

$$= 8\int_0^{\frac{\pi}{4}} \int_0^{\frac{1}{\cos\theta}} (r^2-r^3\cos\theta-r^3\sin\theta+r^4\sin\theta\cos\theta)\, dr\, d\theta$$

$$= 8\int_0^{\frac{\pi}{4}} \left(\frac{\sec^3\theta}{12}-\frac{\sec^3\theta\tan\theta}{20}\right) d\theta$$

$$= 8\left(\frac{\sec\theta\tan\theta+\ln|\sec\theta+\tan\theta|}{24}-\frac{\sec^3\theta}{60}\right)\Bigg|_0^{\frac{\pi}{4}}$$

$$= 8\left(\frac{\sqrt{2}(1)+\ln|\sqrt{2}+1|}{24}-\frac{2\sqrt{2}}{60}+\frac{1}{60}\right)$$

$$= \frac{2+\sqrt{2}+5\ln(\sqrt{2}+1)}{15}$$

$$\approx 0.5214$$

Therefore, $\Delta(2) = 0.5214$.

□

Now move on to higher dimensions. We could attempt to calculate $\Delta(n)$ the same way as in Problem 2. The exact value of $\Delta(n)$ is given by the expression:

$$\Delta(\mathrm{n}) = 2^n \int_0^1 \dots \int_0^1 (1-d_1)(1-d_2)\dots(1-d_n)\sqrt{{d_1}^2 + {d_2}^2 + \dots + {d_n}^2}\,\mathrm{d}d_1\mathrm{d}d_2\dots\mathrm{d}d_n$$

This integral is extremely complicated. Instead of analytically evaluate $\Delta(\mathrm{n})$ for larger n, we will try to approximate $\Delta(\mathrm{n})$, and find whether $\Delta(\mathrm{n})$ converges as n goes to infinity.

The following corollary will be useful to approximating $\Delta(\mathrm{n})$.

Corollary:

If $D_1, \dots, D_n$ are independent random variables with finite positive mean μ and variance σ^2, $\mathrm{Y} = \sum_{i=1}^n D_i$, and $\mathrm{W} = \sqrt{|Y|}$, then $\mathrm{W} \sim \mathrm{N}\left(\sqrt{n\mu - \frac{\sigma^2}{4\mu}}\,,\, \frac{\sigma^2}{4\mu}\right)$ as $\mathrm{n} \to \infty$.

Proof:

Suppose that $D_1, \dots, D_n$ are independent random variables with $\mu > 0$ and variance σ^2 , and $\mathrm{Y} = \sum_{i=1}^n D_i$. Pick some $\alpha \in \mathbb{R}$. By the Central Limit Theorem, $\mathrm{P}\left(\frac{\mathrm{Y}-\mathrm{n}\mu}{\sigma\sqrt{n}} \le \alpha\right) \to F_z(\alpha)$, where F_z is the standard normal cumulative distribution function.

Since $\frac{\alpha^2\sigma^2}{4\mu\sigma\sqrt{n}} \to 0^+$ as $\mathrm{n} \to \infty$, and $\mu \neq 0$,

$$\mathrm{P}\left(\frac{\mathrm{Y} - \mathrm{n}\mu}{\sigma\sqrt{n}} \le \alpha + \frac{\alpha^2\sigma^2}{4\mu\sigma\sqrt{n}}\right) \to F_z(\alpha)$$

$$\mathrm{P}\left(\mathrm{Y} \le \left(\frac{\alpha\sigma}{2\sqrt{\mu}} + \sqrt{n\mu}\right)^2\right) \to F_z(\alpha)$$

Noting that $\mu > 0$ implies that $\mathrm{P}(\mathrm{Y} < 0) \to 0$, then

$$\mathrm{P}\left(\sqrt{|Y|} \le \frac{\alpha\sigma}{2\sqrt{\mu}} + \sqrt{n\mu}\right) \to F_z(\alpha)$$

$$\mathrm{P}\left(\frac{\sqrt{|Y|} - \sqrt{n\mu}}{\frac{\sigma}{2\sqrt{\mu}}} \le \alpha\right) \to F_z(\alpha)$$

Hence, $\mathrm{W} = \sqrt{|Y|} \sim N\left(\sqrt{n\mu}, \frac{\sigma}{2\sqrt{\mu}}\right)$ as $\mathrm{n} \to \infty$.

We could use $\sqrt{n\mu}$ to approximate $\mathbb{E}(\mathrm{W})$. However, there is a better approximation.

Since $\left(\sqrt{|Y|}\right) = \sqrt{\mathbb{E}(|\mathrm{Y}|) - \mathrm{Var}\left(\sqrt{|Y|}\right)}$, $\mathbb{E}(|\mathrm{Y}|) \to n\mu$, and $\mathrm{Var}\left(\sqrt{|Y|}\right) \to \frac{\sigma^2}{4\mu}$, we obtain that $(\mathrm{W}) = \mathbb{E}\left(\sqrt{|Y|}\right) \to \sqrt{n\mu - \frac{\sigma^2}{4\mu}}$.

Therefore, $\mathrm{W} \sim \mathrm{N}\left(\sqrt{n\mu - \frac{\sigma^2}{4\mu}}\,,\, \frac{\sigma^2}{4\mu}\right)$ as $\mathrm{n} \to \infty$.

Problem 3:

Find an expression that approximates $\Delta(n)$ for large n. Does $\Delta(n)$ converge, or does it approach infinity?

Solution:

We know from Problem 1 that the random variable D has the density function:

$$f_D(d) = \begin{cases} 2 - 2d & 0 \le d \le 1 \\ 0 & otherwise \end{cases}.$$

Then, the square of the distance D^2 has the density function:

$$f_{D^2}(d) = f_D(\sqrt{d})\frac{d\sqrt{d}}{dd} = \begin{cases} \frac{1}{\sqrt{d}} - 1 & 0 \le d \le 1 \\ 0 & otherwise \end{cases}.$$

This gives a mean of $\frac{1}{6}$, and a variance of $\frac{7}{180}$.

Now in n dimension, the distance of the two points can be represented by $W = \sqrt{{D_1}^2 + {D_2}^2 + \cdots + {D_n}^2}$ where ${D_i}^2$ has the same distribution as D^2 for each i.

By the corollary, $W \sim N\left(\sqrt{n\mu - \frac{\sigma^2}{4\mu}}, \frac{\sigma^2}{4\mu}\right)$, in this case, with $\mu = \frac{1}{6}$, and $\sigma^2 = \frac{7}{180}$. Thus, $W \sim N\left(\sqrt{\frac{n}{6} - \frac{7}{120}}, \frac{7}{120}\right)$, giving an approximate mean for the distance of $\sqrt{\frac{n}{6} - \frac{7}{120}}$, and an approximate variance of $\frac{7}{120}$ when n is large.

Since $\sqrt{\frac{n}{6} - \frac{7}{120}} \to \infty$ as $n \to \infty$, then $W \to \infty$ as $n \to \infty$. Therefore, random points are farther apart in higher dimensions.

Check how good this approximation is with MATLAB.

```
s = 10000000;    % n samples.
   fprintf('        MATLAB  Apprx \n')
   for n = 1:10  % d dimensions.
      delta = mean(sqrt(sum((rand(s,n)-rand(s,n)).^2,2)));
      apprx = sqrt(n/6-7/120);
      fprintf('%6d %7.4f %7.4f\n',n,delta, apprx)
   end
```

The above MATLAB commands produce the following table. The left column is

the result of 10^7 trials of points-picking. The right column is the result of $\sqrt{\frac{n}{6}-\frac{7}{120}}$.

	MATLAB	Apprx
1	0.3334	0.3291
2	0.5214	0.5244
3	0.6617	0.6646
4	0.7776	0.7800
5	0.8785	0.8803
6	0.9690	0.9704
7	1.0517	1.0528
8	1.1281	1.1292
9	1.1998	1.2007
10	1.2675	1.2682

We can see that $\sqrt{\frac{n}{6}-\frac{7}{120}}$ is a not bad approximation of $\Delta(\mathrm{n})$, even for small n.

References:

https://math.stackexchange.com/questions/1976842/how-is-the-distance-of-two-random-points -in-a-unit-hypercube-distributed

https://stats.stackexchange.com/questions/241504/central-limit-theorem-for-square-roots-of-su ms-of-i-i-d-random-variables

https://blogs.mathworks.com/cleve/2017/09/27/how-far-apart-are-two-random-points-in-a-hyp ercube/