Standardized Euclidean distance

Let us consider measuring the distances between our 30 samples in Exhibit 1.1, using just the three continuous variables pollution, depth and temperature. What would happen if we applied formula (4.4) to measure distance between the last two samples, s29 and s30, for example? Here is the calculation:

$$d_{s29,s30} = \sqrt{(6.0-1.9)^2 + (51-99)^2 + (3.0-2.9)^2} = \sqrt{16.81 + 2304 + 0.01} = \sqrt{2320.82}$$

$$= 48.17$$

The contribution of the second variable "depth" to this calculation is huge – one could say that the distance is practically just the absolute difference in the depth values (equal to |51-99| = 48) with only tiny additional contributions from pollution and temperature. This is the problem of standardization discussed in Chapter 3 – the three variables are on completely different scales of measurement and the larger "depth" values have larger inter-sample differences, so they will dominate in the calculation of Euclidean distances.

Some form of standardization is necessary to balance out the contributions, and the conventional way to do this is to transform the variables so they all have the same variance of 1. At the same time we centre the variables at their means – this centring is not necessary for calculating distance, but it makes the variables all have mean zero and thus easier to compare. The transformation commonly called *standardization* is thus as follows:

standardized value = (original value – mean) / standard deviation (4.5)

The means and standard deviations of the three variables are:

	Pollution	*Depth*	*Temperature*
mean	4.517	74.433	3.057
s.d.	2.141	15.615	0.281

Exhibit 4.4 Standardized values of the three continuous variables of Exhibit 1.1

SITE	*ENVIRONMENTAL VARIABLES*		
NO.	***Pollution***	***Depth***	***Temperature***
s1	0.132	-0.156	1.576
s2	-0.802	0.036	-1.979
s3	0.413	-0.988	-1.268
s4	1.720	-0.668	-0.557
s5	-0.288	-0.860	0.154
s6	-0.895	1.253	1.576
s7	0.039	-1.373	-0.557
s8	0.272	-0.860	0.865
s9	-0.288	-0.412	1.221
s10	2.561	-0.348	-0.201
s11	0.926	-1.116	0.865
s12	-0.335	0.613	0.154
s13	2.281	-1.373	-0.201
s14	0.086	0.549	-1.979
s15	1.020	1.637	-0.913
s16	-0.802	0.613	-0.201
s17	0.880	1.381	0.154
s18	-0.054	-0.028	-0.913
s19	-0.662	0.292	1.932
s20	0.506	-0.092	-0.201
s21	-0.101	-0.988	1.221
s22	-1.222	-1.309	-0.913
s23	-0.989	1.317	-0.557
s24	-0.101	-0.668	-0.201
s25	-1.175	1.445	-0.201
s26	-0.942	0.228	1.221
s27	-1.129	0.677	-0.201
s28	-0.522	1.125	0.865
s29	0.693	-1.501	-0.201
s30	-1.222	1.573	-0.557

leading to the table of standardized values given in Exhibit 4.4. These values are now on comparable standardized scales, in units of standard deviation units with respect to the mean. For example, the value 0.693 would signify 0.693 standard deviations above the mean, and –1.222 would signify 1.222 standard deviations below the mean. The distance calculation thus aggregates squared differences in standard deviation units of each variable. As an example, the distance between the last two sites of Table is:

$$d_{s29,s30} = \sqrt{[0.693-(-1.222)]^2 + [-1.501-1.573]^2 + [-0.201-(-.557)]^2}$$

$$= \sqrt{3.667 + 9.449 + 0.127} = \sqrt{13.243} = 3.639$$

Pollution and temperature have higher contributions than before but depth still plays the largest role in this particular example, even after standardization. But this contribution is justified now, since it does show the biggest standardized difference between the samples. We call this the *standardized Euclidean distance*, meaning that it is the Euclidean distance calculated on standardized data. It will be assumed that standardization refers to the form defined by (4.5), unless specified otherwise.

We can repeat this calculation for all pairs of samples. Since the distance between sample A and sample B will be the same as between sample B and sample A, we can report these distances in a triangular matrix – Exhibit 4.5 shows part of this distance matrix, which contains a total of ½×30×29 = 435 distances.

Exhibit 4.5 Standardized Euclidean distances between the 30 samples, based on the three continuous environmental variables, showing part of the triangular distance matrix.

	s1	s2	s3	s4	s5	s6	· · ·	s24	s25	s26	s27	s28	s29
s2	3.681												
s3	2.977	1.741											
s4	2.708	2.980	1.523										
s5	1.642	2.371	1.591	2.139									
s6	1.744	3.759	3.850	3.884	2.619								
s7	2.458	2.171	0.890	1.823	0.935	3.510							
.	.	.	.	.	.	.	.						
.	.	.	.	.	.	.	.						
.	.	.	.	.	.	.	.						
s25	2.727	2.299	3.095	3.602	2.496	1.810	· · ·	2.371					
s26	1.195	3.209	3.084	3.324	1.658	1.086	· · ·	1.880	1.886				
s27	2.333	1.918	2.507	3.170	1.788	1.884	· · ·	1.692	0.770	1.503			
s28	1.604	3.059	3.145	3.204	2.122	0.813	· · ·	2.128	1.291	1.052	1.307		
s29	2.299	2.785	1.216	1.369	1.224	3.642	· · ·	1.150	3.488	2.772	2.839	3.083	
s30	3.062	2.136	3.121	3.699	2.702	2.182	· · ·	2.531	0.381	2.247	0.969	1.648	3.639

Readers might ask how all this has helped them – why convert a data table with 90 numbers to one that has 435, almost five times more? Were the histograms and scatterplots in Exhibits 1.2 and 1.4 not enough to understand these three variables? This is a good question, but we shall have to leave the answer to the Part 3 of the book, from Chapter 7 onwards, when we describe actual analyses of these distance matrices. At this early stage in the book, we can only ask readers to be patient – try to understand fully the concept of distance which will be the main thread to all the analytical methods to come.

The formula for calculating the distance between the two variables, given three persons scoring on each as shown in Figure 1 is:

Eq. 2 $$d = \sqrt{\sum_{i=1}^{p} (v_{1i} - v_{2i})^2}$$

where the difference between two variables' values is taken, and squared, and summed for p persons (in our example p=3). Only one distance would be computed – between v1 and v2.
Let's do the calculations for finding the Euclidean distances between the three persons, given their scores on two variables. The data are provided in Table 1 below ...

Table 1

	1 Var1	2 Var2
Person 1	20	80
Person 2	30	44
Person 3	90	40

Using equation 1 ...

$$d = \sqrt{\sum_{i=1}^{v} (p_{1i} - p_{2i})^2}$$

For the distance between person **1** and **2**, the calculation is:

$$d = \sqrt{(20-30)^2 + (80-44)^2} = 37.36$$

For the distance between person **1** and **3**, the calculation is:

$$d = \sqrt{(20-90)^2 + (80-40)^2} = 80.62$$

For the distance between person **2** and **3**, the calculation is:

$$d = \sqrt{(30-90)^2 + (44-40)^2} = 60.13$$

Using equation 2, we can also calculate the distance between the two variables …

$$d = \sqrt{\sum_{i=1}^{p} (v_{1i} - v_{2i})^2}$$

$$d = \sqrt{(20-80)^2 + (30-44)^2 + (90-40)^2} = 79.35$$

Equation 1 is used where say we are comparing two "objects" across a range of variables – and trying to determine how "dissimilar" the objects are (the Euclidean distance between the two objects taking into account their magnitudes on the range of variables. These objects might be two person's profiles, a person and a target profile, in fact basically any two vectors taken across the same variables.

Equation 2 is used where we are comparing two variables to one another – given a sample of paired observations on each (as we might with a pearson correlation), In our case above, the sample was three persons.

In both equations, **Raw Euclidean Distance** is being computed.

Normalised Euclidean Distance

The problem with the raw distance coefficient is that it has no obvious bound value for the maximum distance, merely one that says 0 = absolute identity. Its range of values vary from 0 (absolute identity) to some maximum possible discrepancy value which remains unknown until specifically computed. Raw Euclidean distance varies as a function of the magnitudes of the observations. Basically, you don't know from its size whether a coefficient indicates a small or large distance.

If I divided every person's score by 10 in Table 1, and recomputed the euclidean distance between the persons, I would now obtain distance values of 3.736 for person 1 compared to 2, instead of 37.36. Likewise, 8.06 for person 1 and 3, and 6.01 for persons 2 and 3. The raw distance conveys little information about absolute dissimilarity.

So, raw euclidean distance is acceptable only if relative ordering amongst a fixed set of profile attributes is required. But, even here, what does a figure of 37.36 actually convey. If the maximum possible observable distance is 38, then we know that the persons being compared are about as different as they can be. But, if the maximum observable distance is 1000, then suddenly a value of 37.36 seems to indicate a pretty good degree of agreement between two persons.

The fact of the matter is that unless we know the maximum possible values for a euclidean distance, we can do little more than rank dissimilarities, without ever knowing whether any or them are actually similar or not to one another in any absolute sense.

A further problem is that raw Euclidean distance is sensitive to the scaling of each constituent variable. For example, comparing persons across variables whose score ranges are dramatically different. Likewise, when developing a matrix of Euclidean coefficients by comparing multiple variables to one another, and where those variables' magnitude ranges are quite different.

For example, say we have 10 variables and are comparing two person's scores on them … the variable scores might look like …

Table 2

	1 Person 1	2 Person 2
Var 1	1	2
Var 2	1	1
Var 3	4	5
Var4	6	6
Var5	1200	1300
Var6	3	3
Var7	2	2
Var8	3	5
Var9	2	3
Var10	8	8

해당 수치는 '|1200 - 1300|'라 거의 동일함.

The two persons' scores are virtually identical (except for variable 5.) The raw Euclidean distance for these data is: **100.03**. If we had expressed the scores for variable 5 in the same metric as the other scores (on a 1-10 metric scale), we would have scores of 1.2 and 1.3 respectively for each individual. The raw Euclidean distance is now: **2.65**.

Obviously, the question "is 2.65 good or bad" still exists – given we have no idea what the maximum possible Euclidean distance might be for these data.

This is where SYSTAT, Primer 5, and SPSS provide Standardization/Normalization options for the data so as to permit an investigator to compute a distance coefficient which is essentially "scale free".