Processing new types of quality data:
analysis of partitioning metric space data

Yariv N. Marmor
Emil Bashkansky

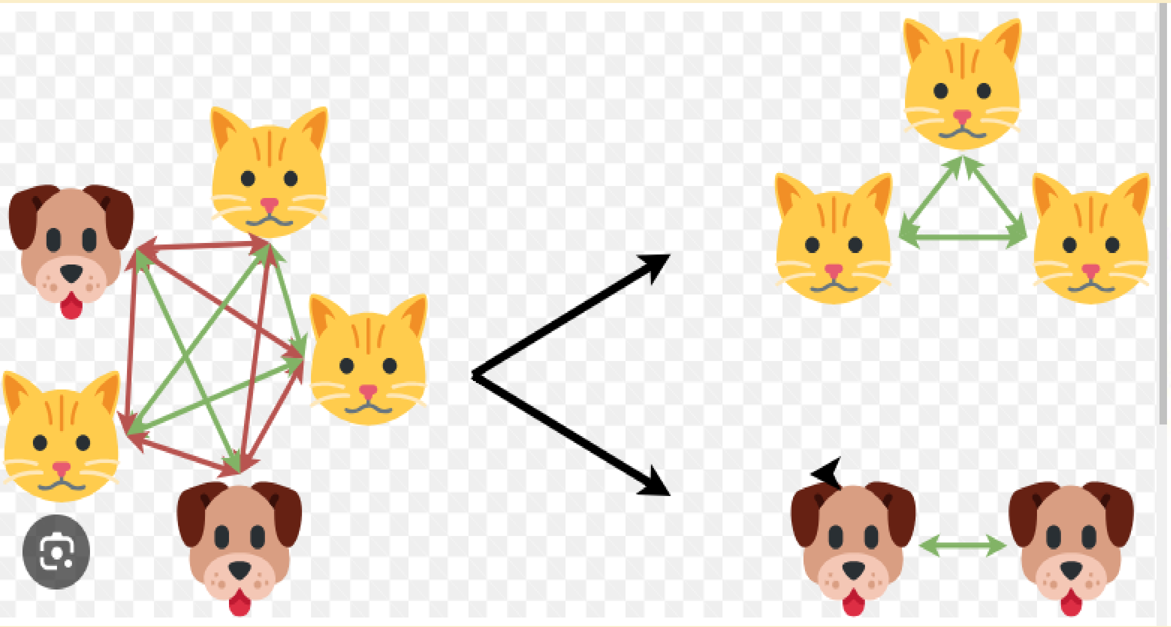
BRAUDE - College of Engineering,
Karmiel, ISRAEL

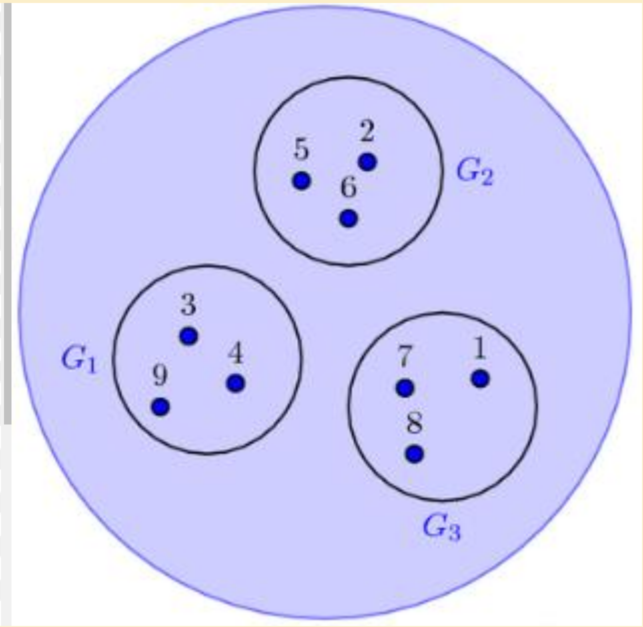
Valencia - ENBIS 2023
September, 13

Yariv N. Marmor, Emil Bashkansky " Processing new types of quality data"
Quality and Reliability Engineering International , 36 (8), 2621-2638, 2020

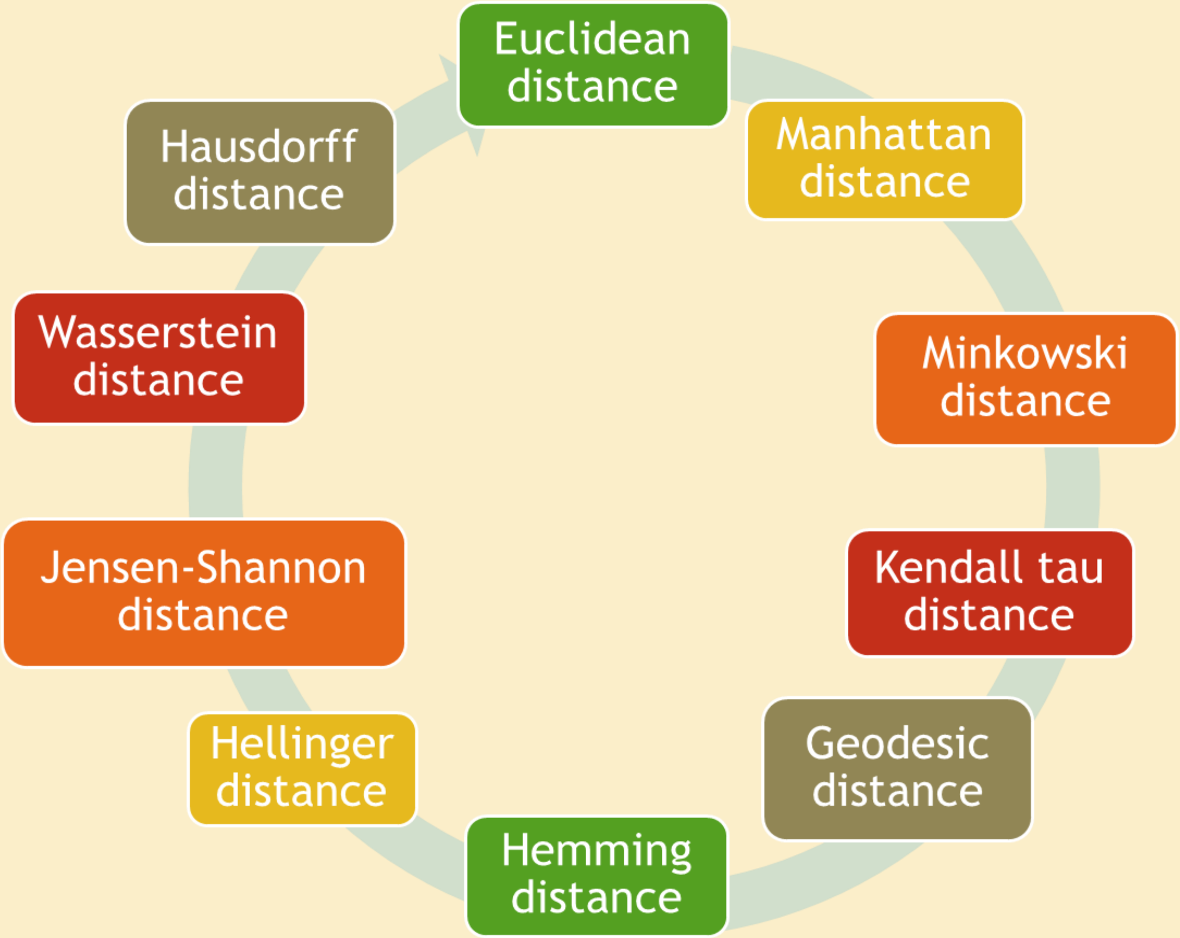
	2.1 Nominal data	2.2 Ordinal data	2.3 Ranked/ prioritized data	2.4 Strings	2.5 Tree-structured data	2.6 Product or process distribution
3.1 Similarity/dissimilarity and closeness/distance	4.1 4.1.1	4.1 4.1.2	4.1 4.1.3	4.1 4.1.4	4.1 4.1.5	4.1 4.1.6
3.2 Metrological aspects	4.2 4.2.1	4.2 4.2.2	4.2 4.2.3	4.2 4.2.4	4.2 4.2.5	4.2 4.2.6
3.3 Data aggregation or fusion and location measure	4.3 4.3.1	4.3 4.3.2	4.3 4.3.3	4.3 4.3.4	4.3 4.3.5	4.3 4.3.6
3.4 Dispersion measure and variation analysis	4.4 4.4.1	4.4 4.4.2	4.4 4.4.3	4.4 4.4.4	4.4 4.4.5	4.4 4.4.6
3.5 Data partitioning			*			*
3.6 Data clustering…						

Data/objects partition





Examples of distance metrics



Euclidean distance

Manhattan distance

Minkowski distance

Kendall tau distance

Geodesic distance

Hemming distance

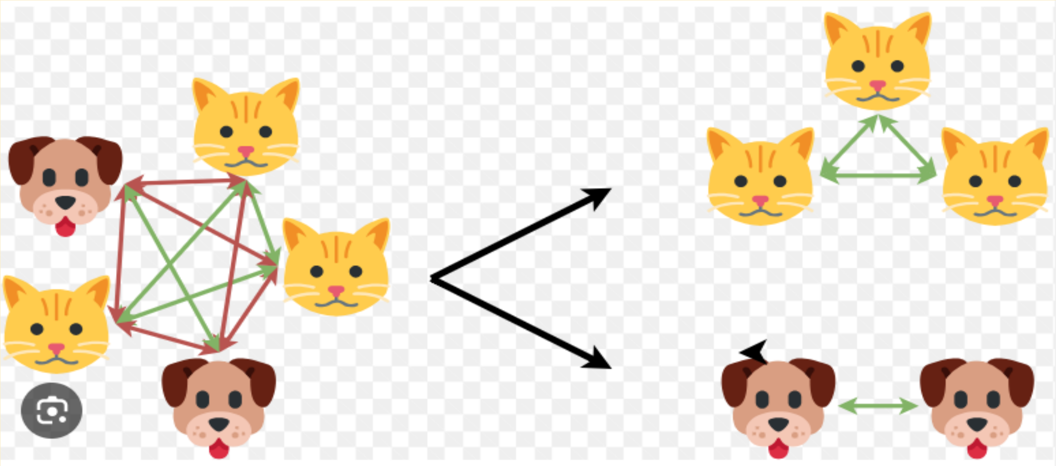
Hellinger distance

Jensen-Shannon distance

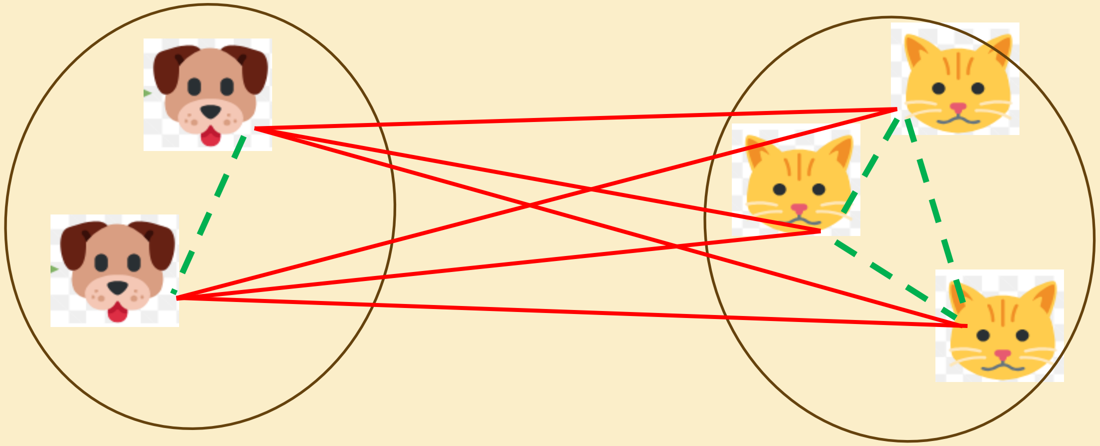
Wasserstein distance

Hausdorff distance

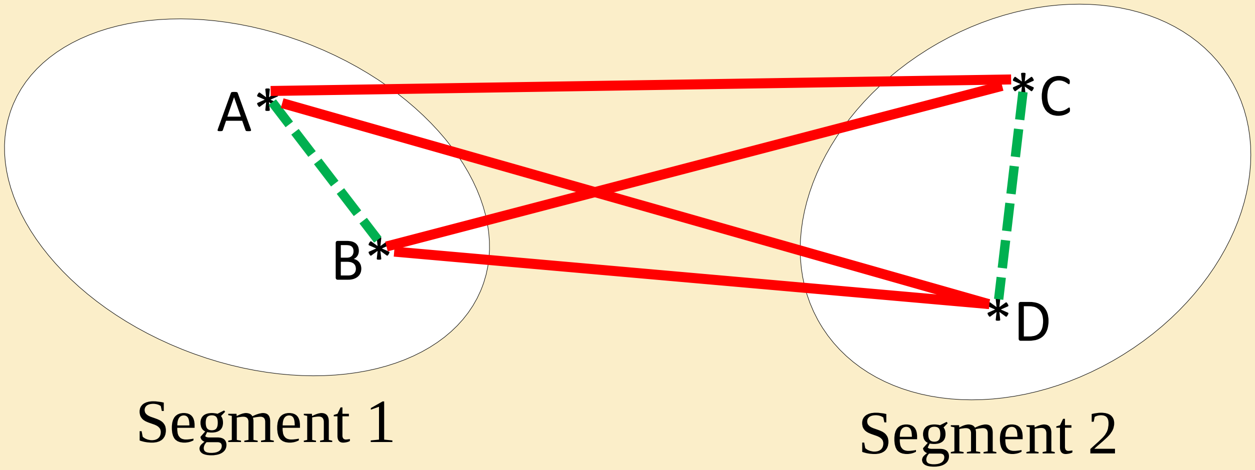
Data/objects partition:
degrees of connection (dc) split



$$dc_{\text{total}} = dc_{\text{inter}} + dc_{\text{intra}}$$



Data/objects partition: *sum of distances (SD) split*



Sum of Distances: Intra $\{(SD)_{intra} = d(A,B)+d(C,D)\}$

Sum of Distances: Inter $\{(SD)_{inter} = d(A,C)+d(A,D)+d(B,C)+d(B,D)\}$

Sum of Distances: Total

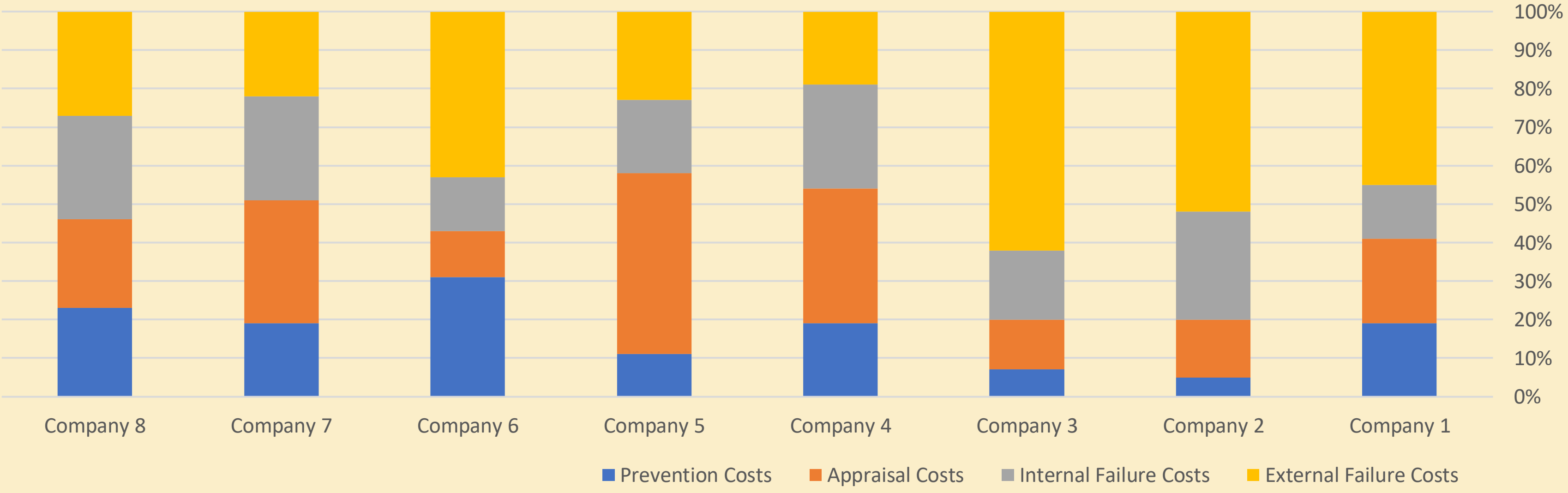
$$\{(SD)_{total} = d(A,B)+d(C,D)+d(A,C)+d(A,D)+d(B,C)+d(B,D)\}$$

Segregation power index - SP

$$SP = \frac{MD_{\text{inter}}}{MD_{\text{intra}}} = \frac{SD_{\text{inter}} / dc_{\text{inter}}}{SD_{\text{intra}} / dc_{\text{intra}}}$$

Example 1: Distribution of quality costs proportions by the four main categories (Rosenfeld et.al, 2019)

	Company 1	Company 2	Company 3	Company 4	Company 5	Company 6	Company 7	Company 8
Prevention Costs	0.19	0.05	0.07	0.19	0.11	0.31	0.19	0.23
Appraisal Costs	0.22	0.15	0.13	0.35	0.47	0.12	0.32	0.23
Internal Failure Costs	0.14	0.28	0.18	0.27	0.19	0.14	0.27	0.27
External Failure Costs	0.45	0.52	0.62	0.19	0.23	0.43	0.22	0.27



Hellinger distances and SP calculation

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{4} \left(\sqrt{p_i} - \sqrt{q_i} \right)^2}$$

	Company 1	Company 2	Company 3	Company 4	Company 5	Company 6	Company 7	Company 8
Company 1	0.000	0.198	0.169	0.214	0.222	0.122	0.189	0.152
Company 2	0.198	0.000	0.094	0.290	0.290	0.265	0.265	0.240
Company 3	0.169	0.094	0.000	0.328	0.320	0.230	0.302	0.266
Company 4	0.214	0.290	0.328	0.000	0.120	0.269	0.030	0.104
Company 5	0.222	0.290	0.320	0.120	0.000	0.317	0.127	0.191
Company 6	0.122	0.265	0.230	0.269	0.317	0.000	0.244	0.178
Company 7	0.189	0.265	0.302	0.030	0.127	0.244	0.000	0.077
Company 8	0.152	0.240	0.266	0.104	0.191	0.178	0.077	0.000

$SD_{intra} = 1.727$	$dc_{intra} = 12$	$MD_{intra} = SD_{intra}/12 = 0.144$
$SD_{inter} = 4.082$	$dc_{inter} = 16$	$MD_{inter} = SD_{inter}/16 = 0.255$
$SD_{total} = 5.809$	$dc_{total} = 28$	SP = 1.773

Examples of 2 preference chains consisting of 4 alternatives

$$C_1 : A_3 > A_4 > A_1 > A_2$$

$$C_2 : A_2 > A_1 > A_4 > A_3$$

Example 2: Preference/prioritization chains

$$C_1 : A_3 > A_4 > A_1 > A_2$$

$$C_2 : A_2 > A_1 > A_4 > A_3$$

	A_1	A_2	A_3	A_4
A_1	0	1	-1	-1
A_2	-1	0	-1	-1
A_3	1	1	0	1
A_4	1	1	-1	0

	A_1	A_2	A_3	A_4
A_1	0	-1	1	1
A_2	1	0	1	1
A_3	-1	-1	0	-1
A_4	-1	-1	1	0

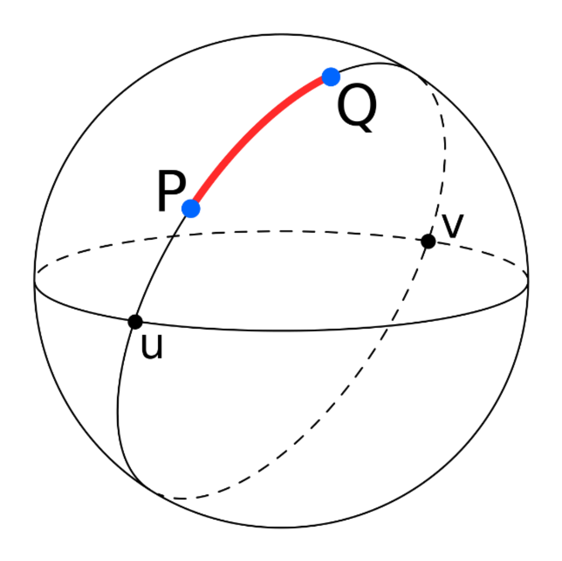
C_1

(0 1 -1 -1 -1 0 -1 -1 1 1 0 1 1 1 -1 0)

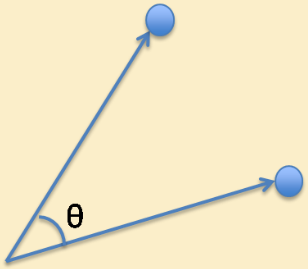
C_2

(0 -1 1 1 1 0 1 1 -1 -1 0 -1 -1 -1 1 0)

Distance metric
based on cosine similarity



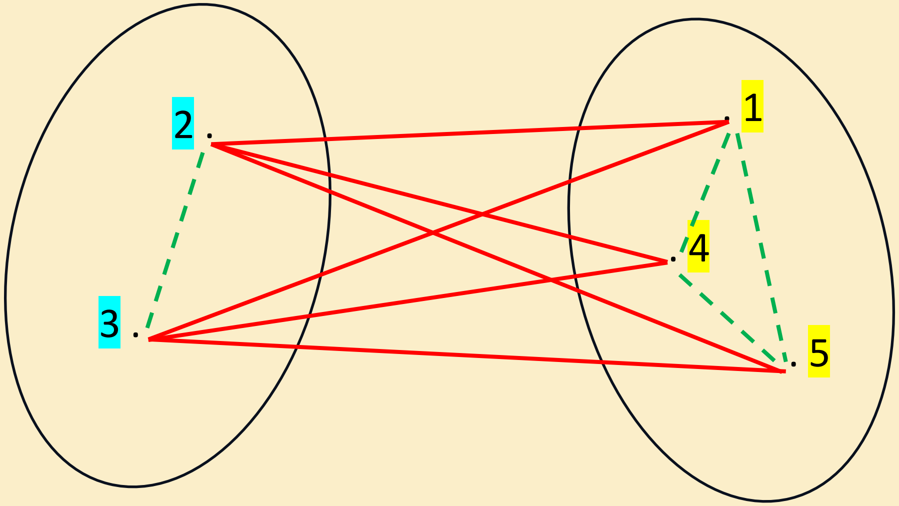
$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$



$$d\,(\vec{a},\vec{b}) = \frac{\hat{\theta}_{\vec{a},\vec{b}}}{\pi} = \frac{\arccos\left(\frac{\sum_{\forall i} a_i \cdot b_i}{\sqrt{\sum_{\forall i} a_i^2} \cdot \sqrt{\sum_{\forall i} b_i^2}}\right)}{\pi}$$

Five experts/judges have prioritized alternatives with judges no. 2 and no. 3 being women, while judges no.1, no.4 and no.5 were men (Vanacore et.al, 2019)

		Judge *j*				
		1	*2*	*3*	*4*	*5*
Judge *i*	*1*	0	0.59	0.73	0.33	0.38
	2	0.59	0	0.46	0.45	0.44
	3	0.73	0.46	0	0.65	0.61
	4	0.33	0.45	0.65	0	0.37
	5	0.38	0.44	0.61	0.37	0



$$MD_{\ intra} = [d_{2,3} + (d_{1,4} + d_{1,5} + d_{4,5})]/4 = 0.385$$

$$MD_{\ inter} = [(d_{2,1} + d_{2,4} + d_{2,5}) + (d_{3,1} + d_{3,4} + d_{3,5})]/6 = 0.578$$

$$\mathbf{SP} = MD_{\ inter} \,/\, MD_{\ intra} = \mathbf{1.502}$$

SP distribution under given null/ homogeneity hypothesis H_0

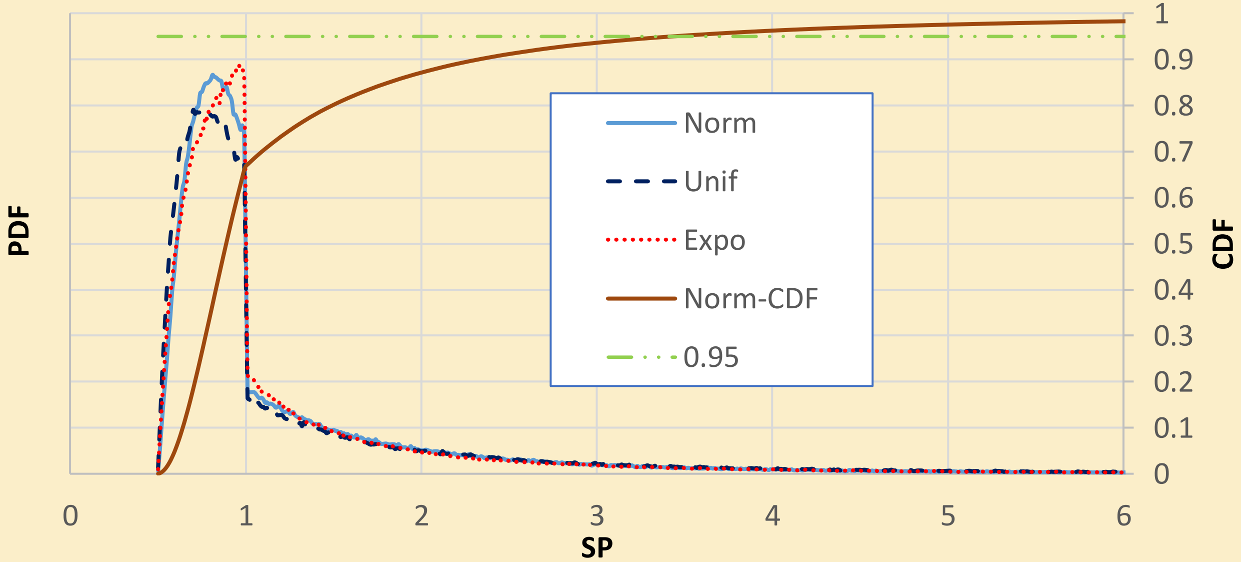
Does not depend on the location parameter of the origin data distribution

Does not depend on the scale parameter of the origin data distribution

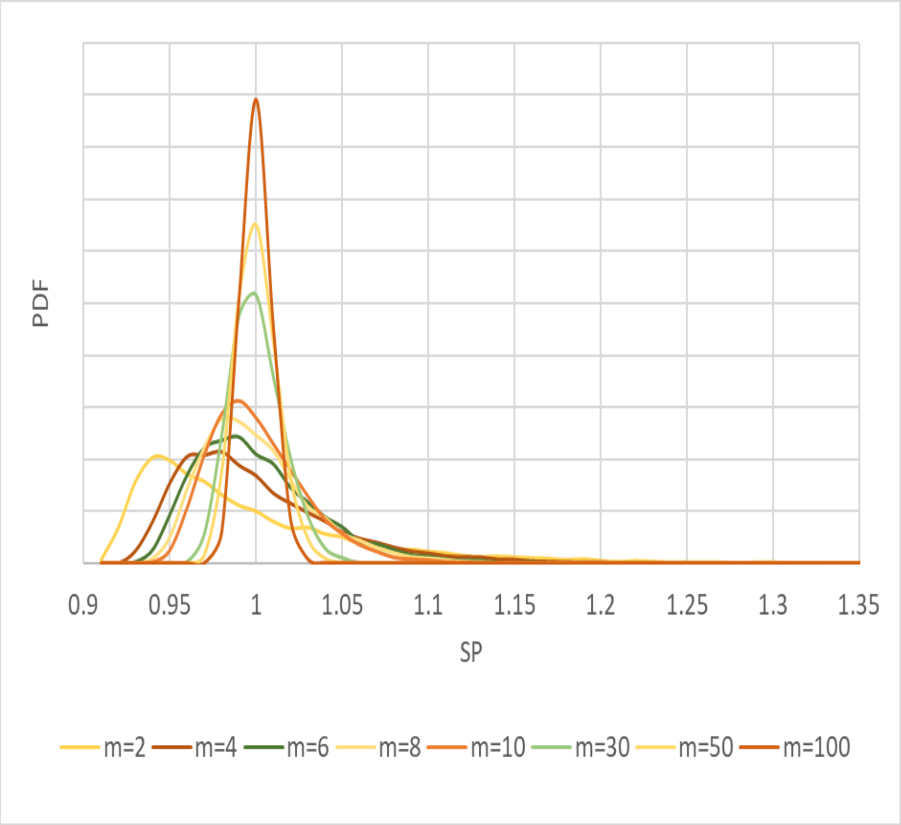
Almost independent on shape parameter, especially for *SP* >1

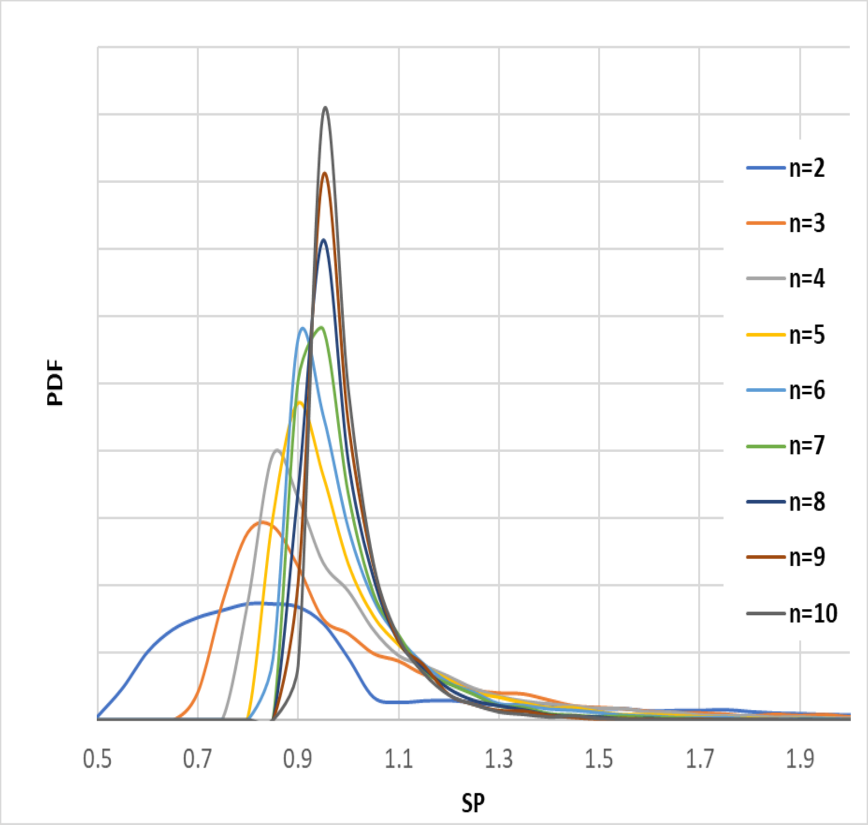
Depends only on the type of partition, i.e., vector $(n_1, n_2, \ldots, n_{k,} \ldots, n_m)$

(2,2)



How amount of data influence SP distribution under H_0





$(10,10,\ldots 10)$

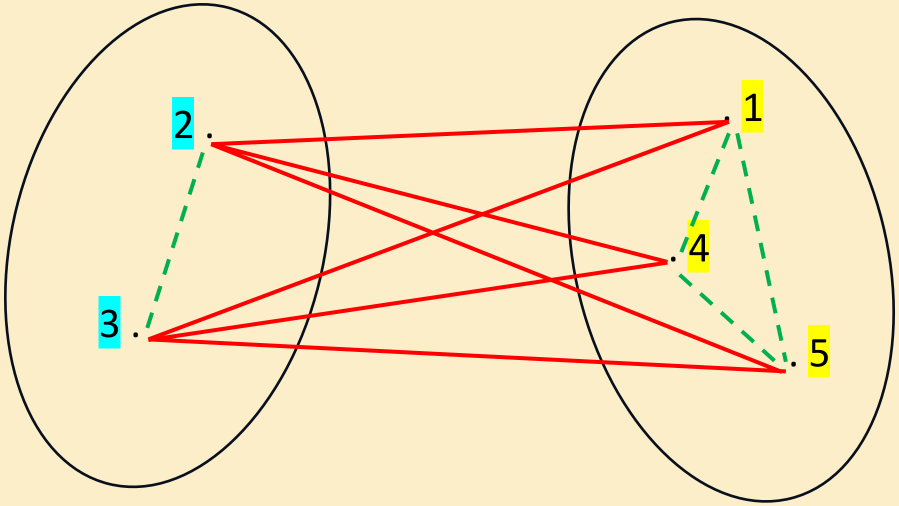
(n, n)

$n = 10$

$m = 2$

Five experts/judges have prioritized alternatives with judges no. 2 and no. 3 being women, while judges no.1, no.4 and no.5 were men (Vanacore et.al, 2019)

		Judge *j*				
		1	*2*	*3*	*4*	*5*
Judge *i*	*1*	0	0.59	0.73	0.33	0.38
	2	0.59	0	0.46	0.45	0.44
	3	0.73	0.46	0	0.65	0.61
	4	0.33	0.45	0.65	0	0.37
	5	0.38	0.44	0.61	0.37	0



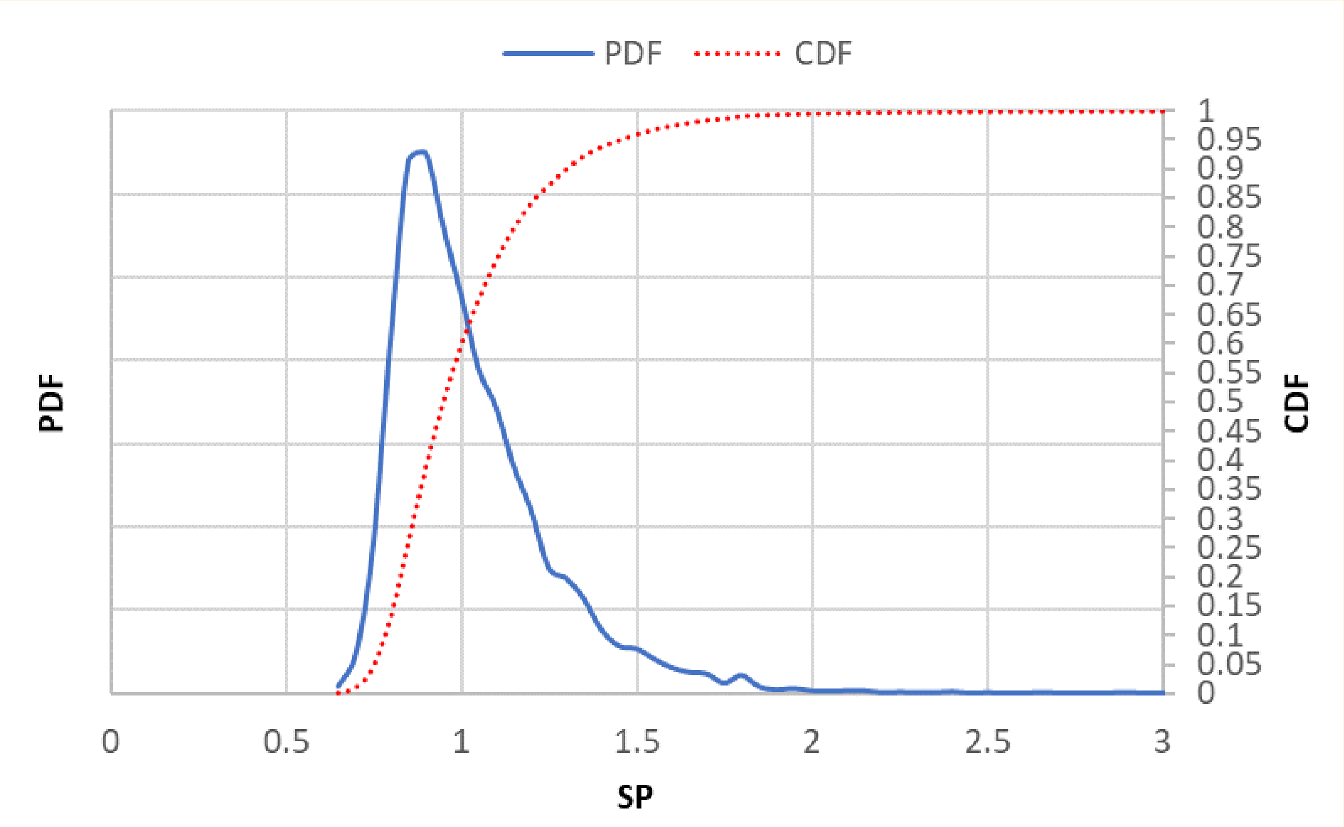
$$MD_{intra} = [d_{2,3} + (d_{1,4} + d_{1,5} + d_{4,5})]/4 = 0.385$$

$$MD_{inter} = [(d_{2,1} + d_{2,4} + d_{2,5}) + (d_{3,1} + d_{3,4} + d_{3,5})]/6 = 0.578$$

$$\mathbf{SP} = MD_{inter} / MD_{intra} = \mathbf{1.502}$$

The null hypothesis:

“Gender equality and the absence of real preferences between alternatives”



$SP_{0.95} = 1.476$ and p – value for calculated SP =1.502 equals 4.39%

Modus operandi – 10 steps

1. Decide on the population of objects under study (OUS).
2. Make assumption about the type of the expected distribution of these objects within a homogeneous population.
3. Choose distance metric suitable to this distribution
4. Decide on the criterion that, in your opinion, can influence (heterogeneity hypothesis) and which levels serve as the basis for dividing/separating objects into groups (partition).
5. Provide corresponding data partitioning/division
6. Calculate *SP*
7. Simulate *SP* distribution under H_0 in accordance with partition vector ($n_1, n_2, \ldots, n_{k,} \ldots, n_m$) and the chosen distance metric. Every cycle of simulation process includes:
    a. Random generation of *N* data from a population of OUS (in accordance to step 1) characterized by the assumed distribution (in accordance to step 2).
    b. Distance matrix calculation (according to step 3)
    c. Partitioning these distances to *inter* and *intra* components in accordance to steps 4 and 5
    d. SP calculation and back to a.
8. Determine the alpha risk - α of homogeneity hypothesis H_0 rejection.
9. Find (1-α) percentile of simulated *SP* distribution, or alternatively, *p* – value of calculated *SP*.
10. Make final conclusion about the expediency of the made partition and its discrimination/separation/segregation power by usual statistical methods according to previous step results.

Thank You for your attention!

E-mail: ebashkan@braude.ac.il