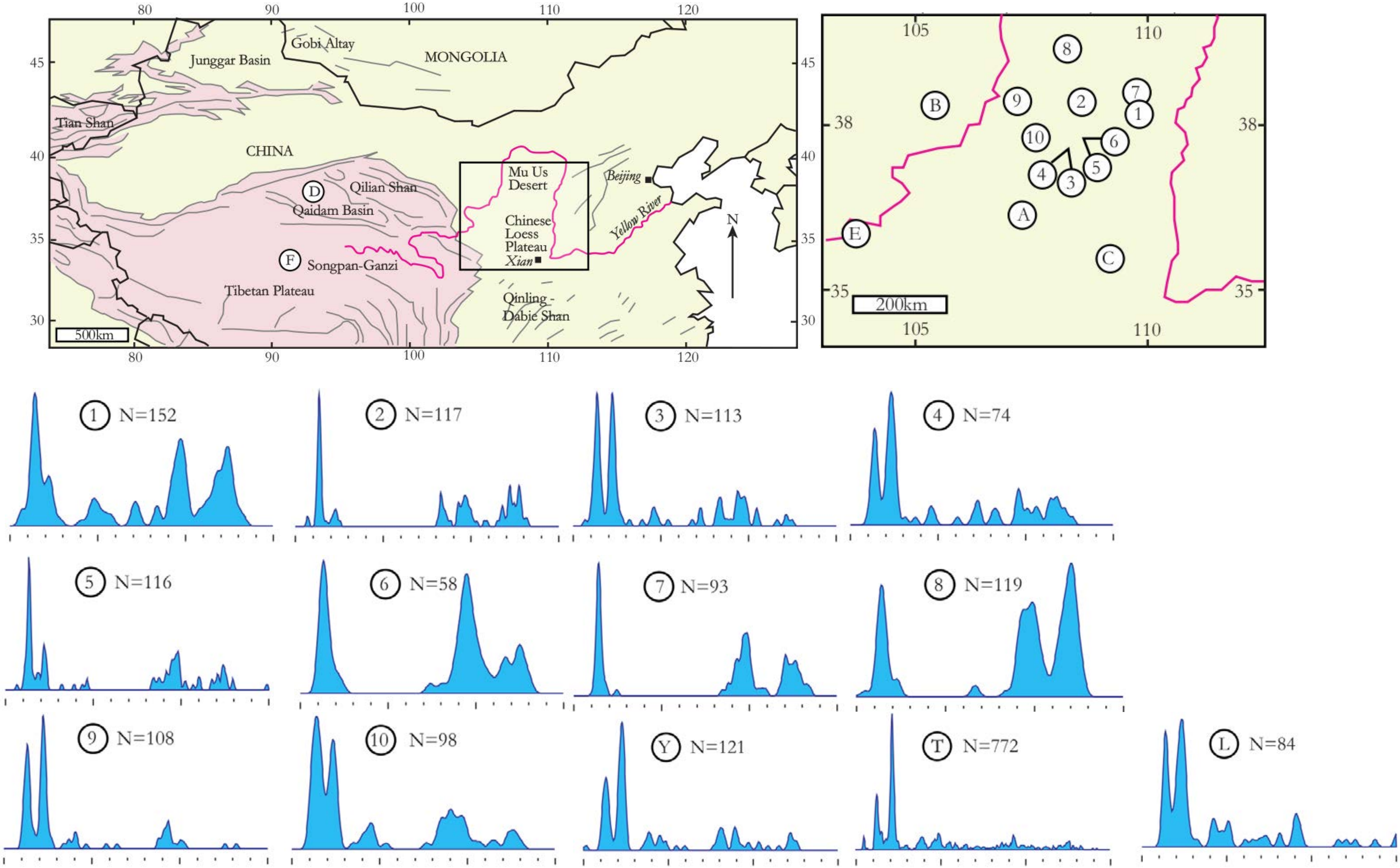
ON THE VISUALISATION AND INTERCOMPARISON OF DETRITAL AGE DISTRIBUTIONS

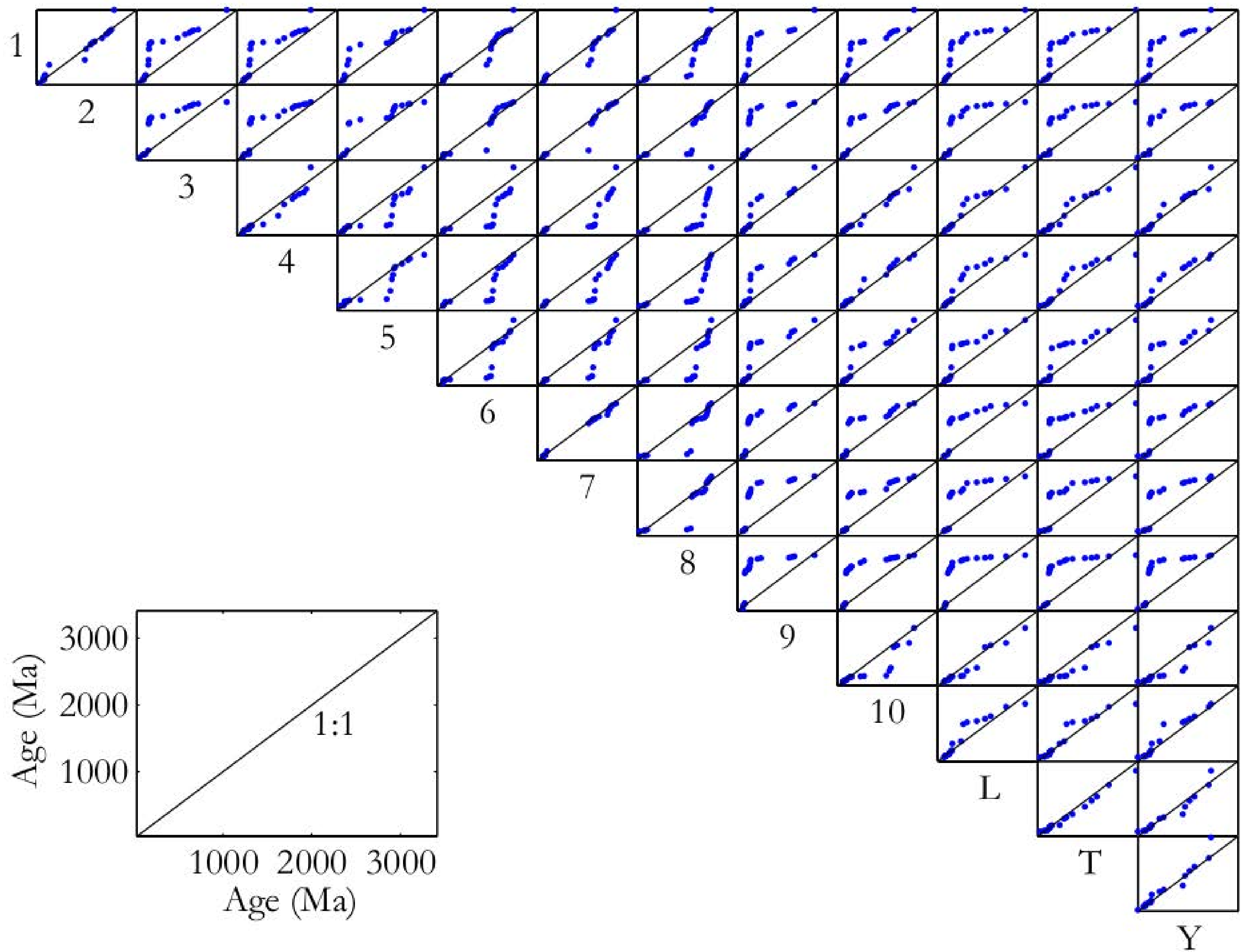
PART 2: INTERCOMPARISON

Pieter Vermeesch

London Geochronology Centre
University College London
p.vermeesch@ucl.ac.uk

A Chinese case study with 2,025 ages from 13 samples





n(n-1)/2 = 78 pairwise comparisons

A ‘statistic’ is “*any quantity whose value can be calculated from sample data*”.

(Devore, 2011)

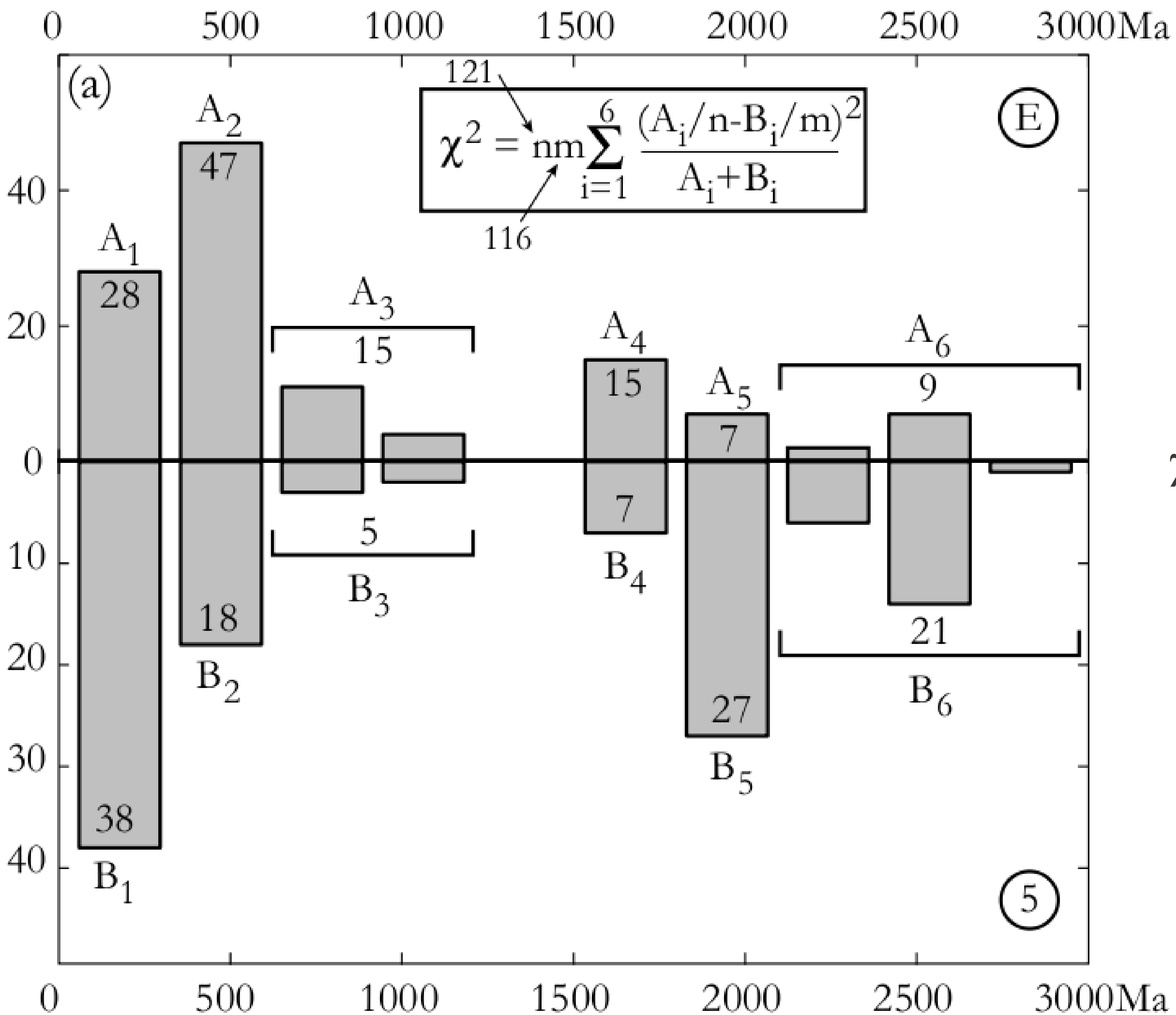
Examples:

$$\rightarrow \textit{arithmetic mean}: \bar{x} = \sum_{i=1}^{N} x_i / N$$

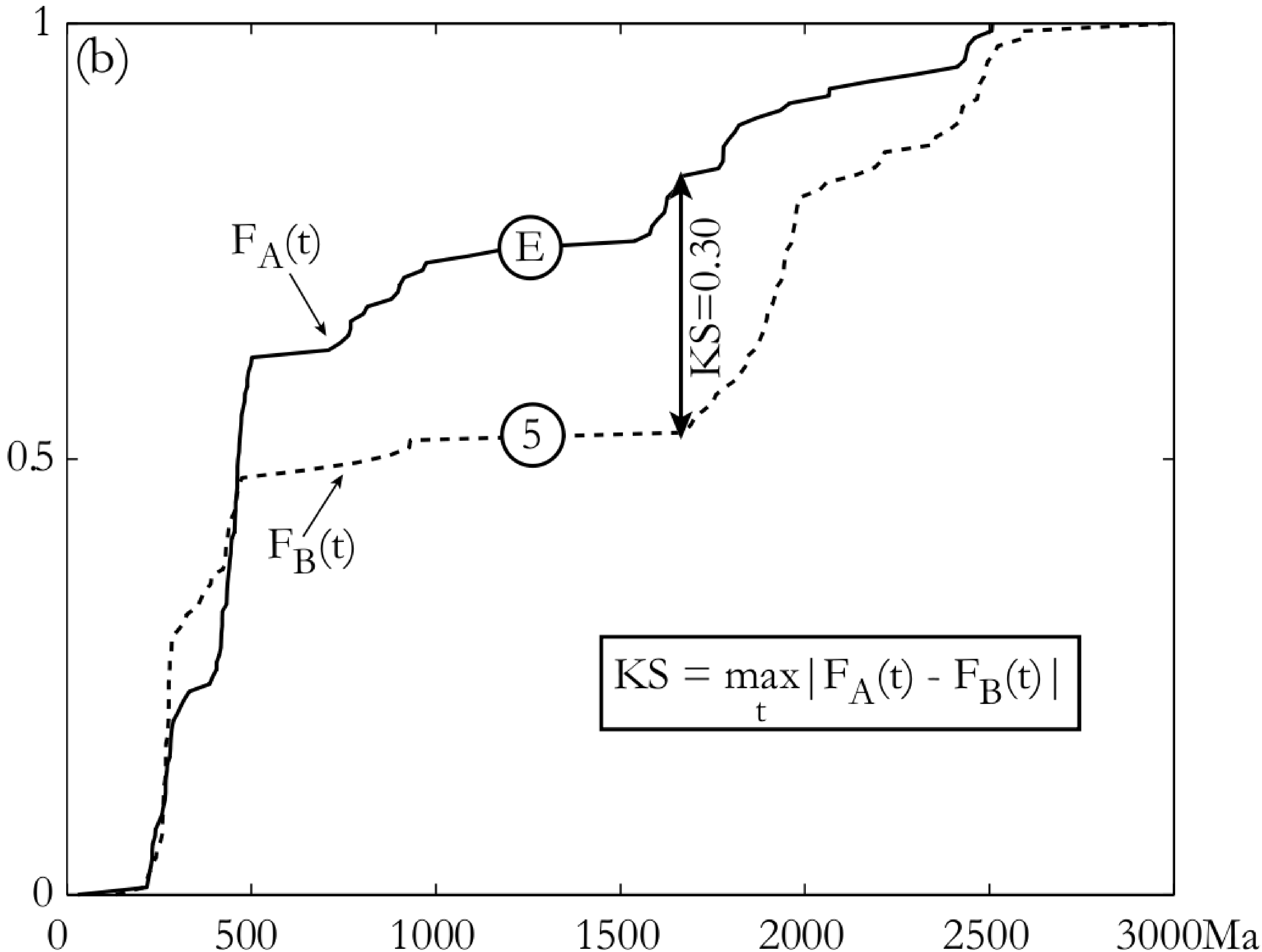
$$\rightarrow \textit{standard deviation}: \sigma(x) = \sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 / (N-1)}$$

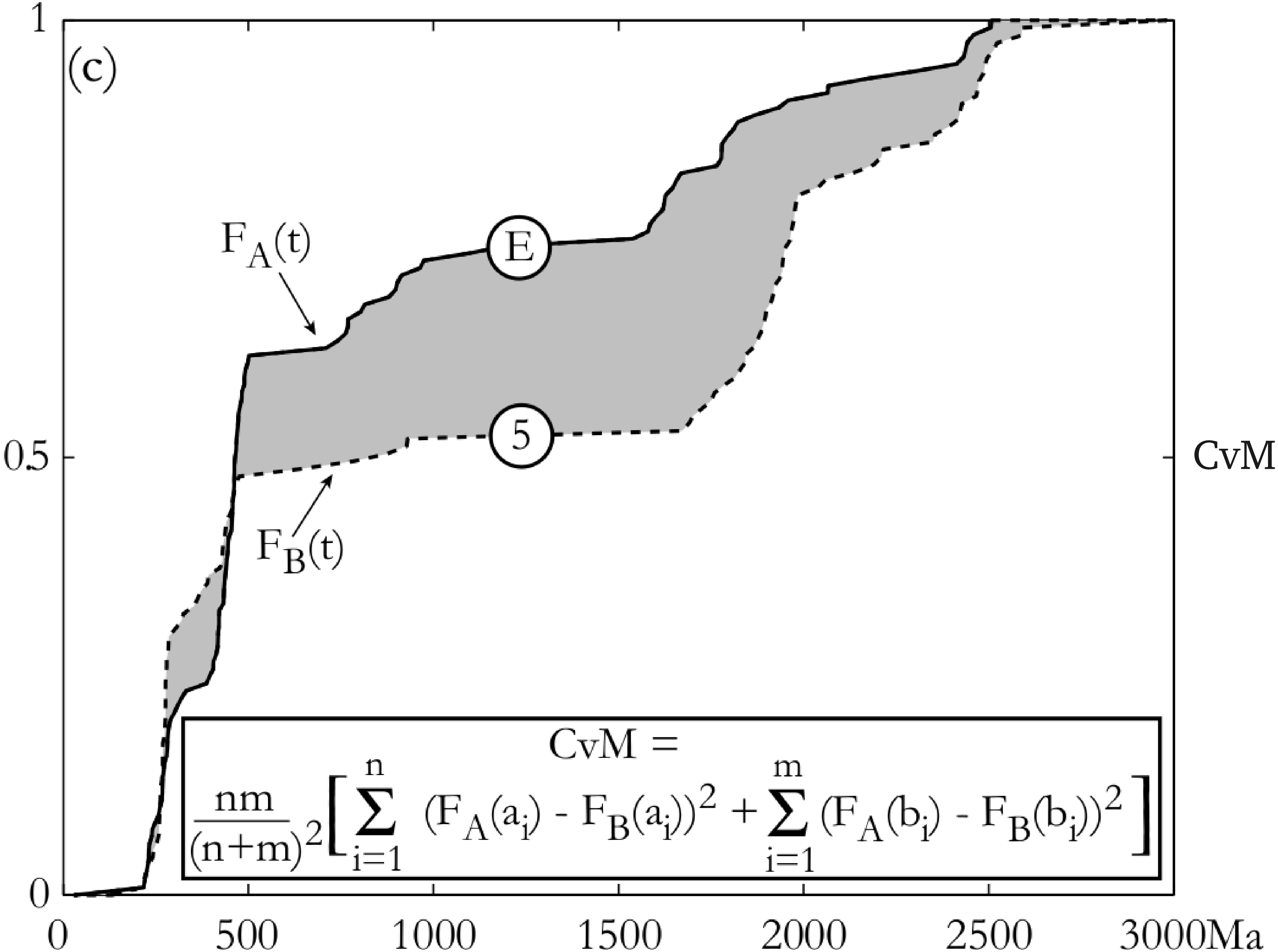
$$\rightarrow \textit{maximum}: M = max\{x_i\}$$

...


0
500
1000
1500
2000
2500
3000Ma
(a)
121
$\chi^2 = nm\sum_{i=1}^{6}\frac{(A_i/n-B_i/m)^2}{A_i+B_i}$
116
E
A_1
28
A_2
47
A_3
15
A_4
15
A_5
7
A_6
9
40
20
0
10
20
30
40
5
7
18
38
27
21
B_1
B_2
B_3
B_4
B_5
B_6
5

$\chi^2 = 38.8$


(b)
1
0.5
0
0
500
1000
1500
2000
2500
3000Ma
$F_A(t)$
$F_B(t)$
E
5
KS=0.30
$KS = \max_t |F_A(t) - F_B(t)|$

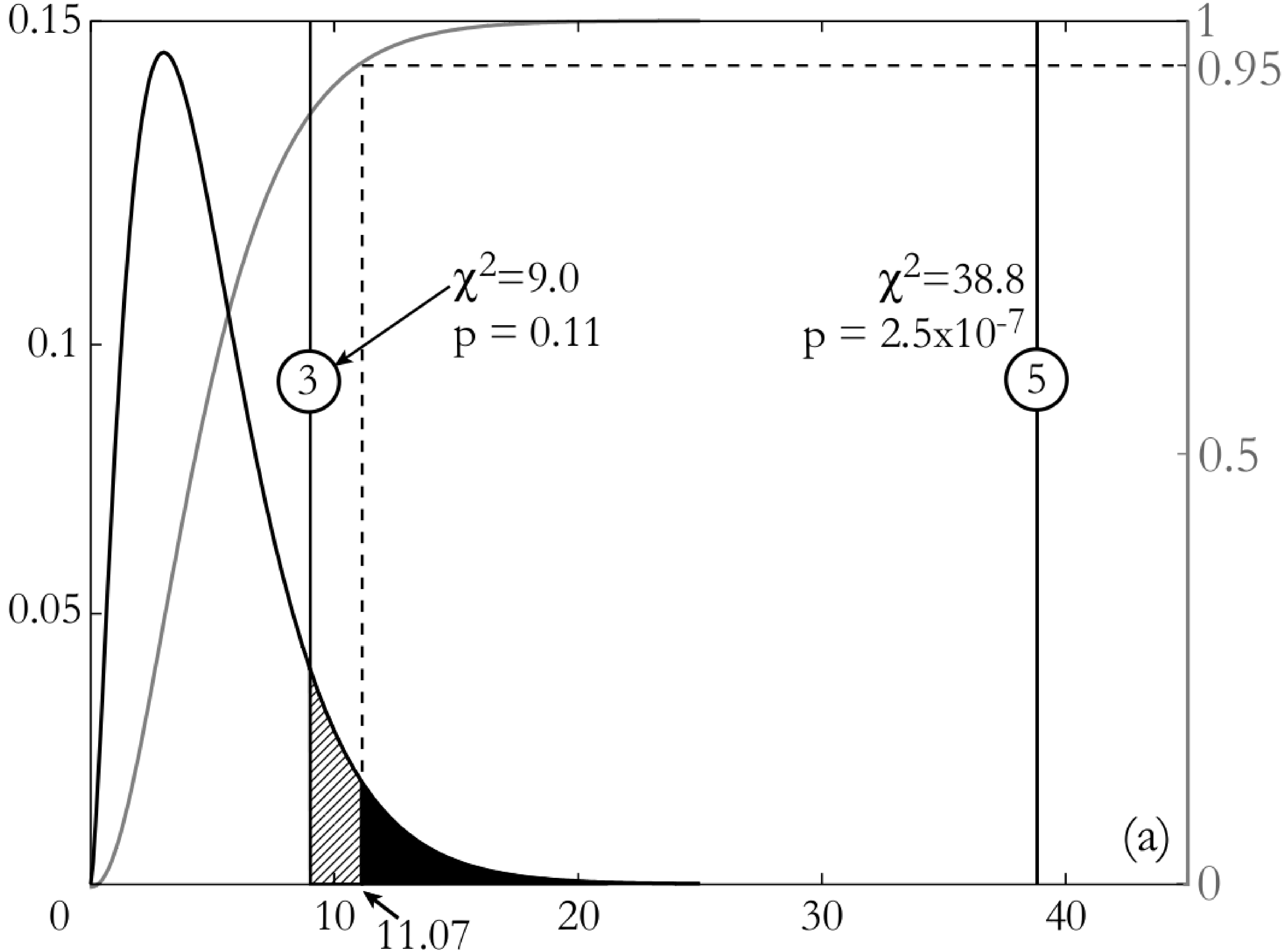

(c)
$F_A(t)$
$F_B(t)$
E
5
CvM =
$$\frac{nm}{(n+m)^2}\left[\sum_{i=1}^{n}(F_A(a_i) - F_B(a_i))^2 + \sum_{i=1}^{m}(F_A(b_i) - F_B(b_i))^2\right]$$
1
0.5
0
0
500
1000
1500
2000
2500
3000Ma

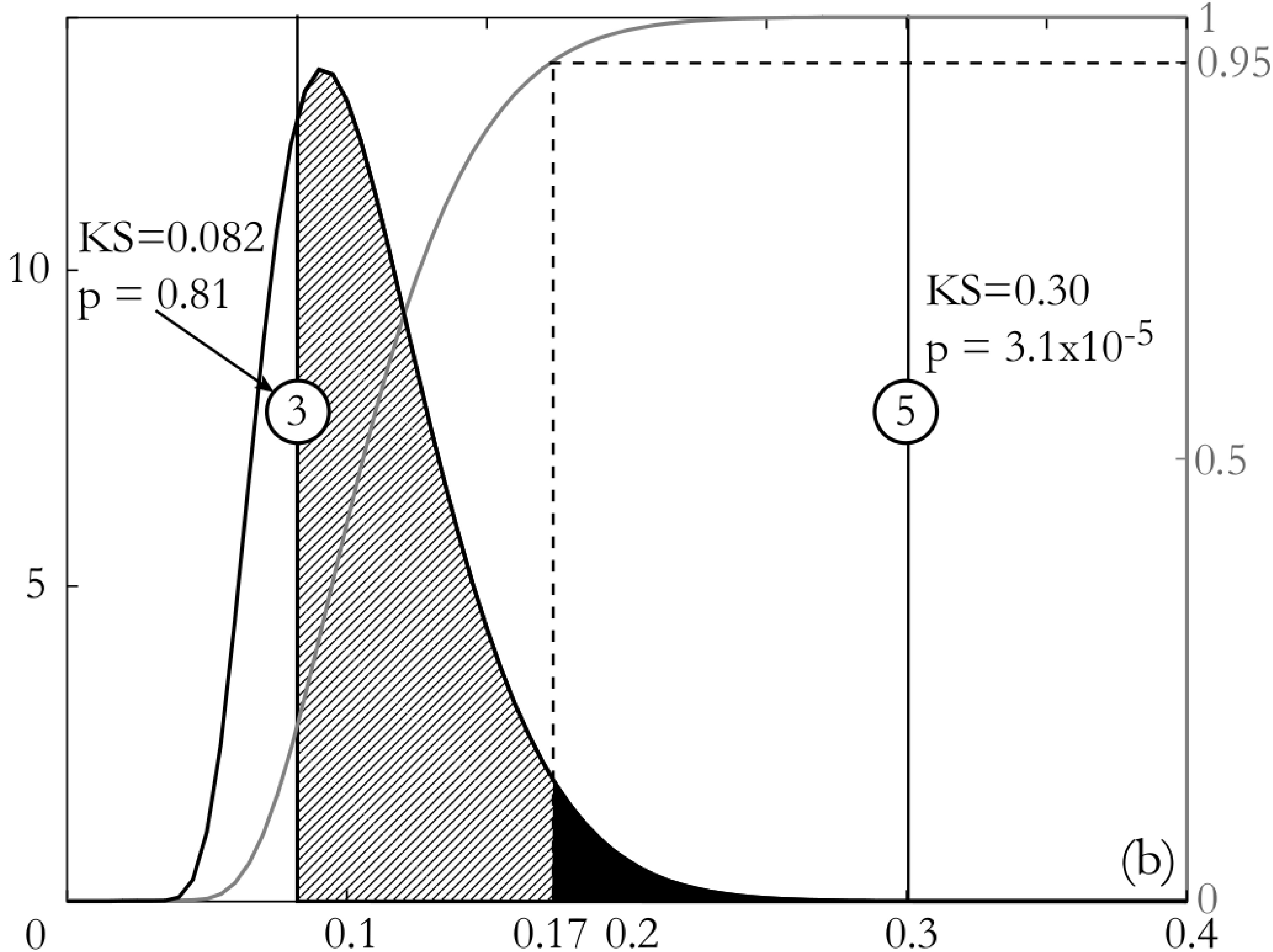
CvM = 1.07

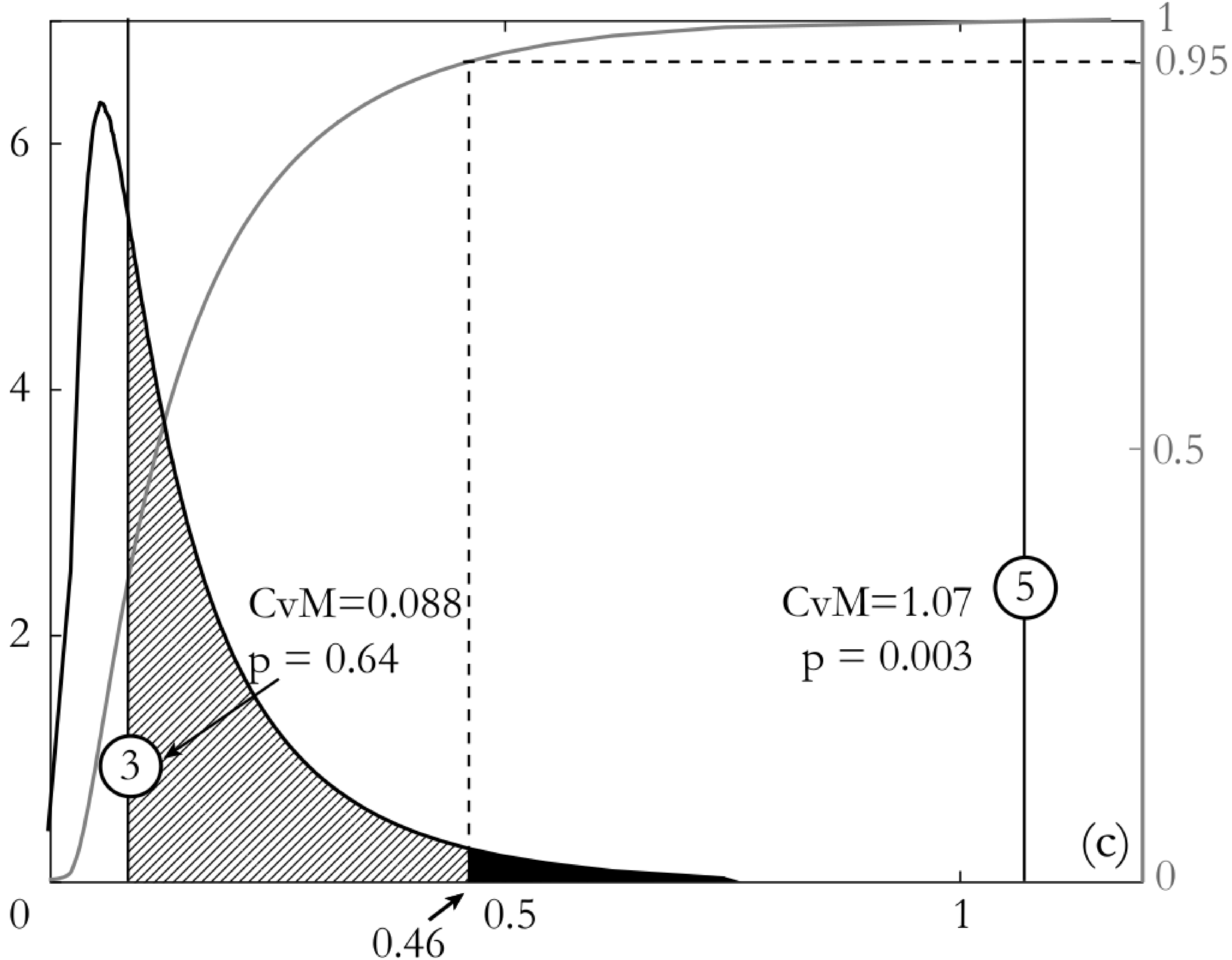
1. formulate a 'null hypothesis' and an 'alternative hypothesis'
   e.g. H_o: "*Two samples were drawn from the same distribution*"
   H_a: "*Two samples were drawn from different distributions*"

2. given a dataset D, calculate the 'test statistic' S(D)

3. if S(D) is 'unlikely' under H_o, then abandon the latter in favour of H_a


0.15
0.1
0.05
0
10
20
30
40
11.07
1
0.95
0.5
0
χ²=9.0
p = 0.11
χ²=38.8
p = 2.5x10⁻⁷
3
5
(a)


KS=0.082
p = 0.81
3
KS=0.30
p = 3.1x10-5
5
10
5
0
0.1
0.17
0.2
0.3
0.4
1
0.95
0.5
0
(b)


6
4
2
0
0.5
1
0.46
1
0.95
0.5
0
CvM=0.088
p = 0.64
3
CvM=1.07
p = 0.003
5
(c)

Three factors determine the outcome of a statistical test:

1. The significance criterion (α)

2. The sample size (n)

3. The effect size (ε)

With the effect size being:

"*the degree to which the null hypothesis is false*"

(Cohen, 1977).

61,850 citations can't be wrong!

Google | jacob cohen statistical power analysis for the behavioral sciences

Scholar — About 12,200 results (**0.07** sec)

Articles
Legal documents

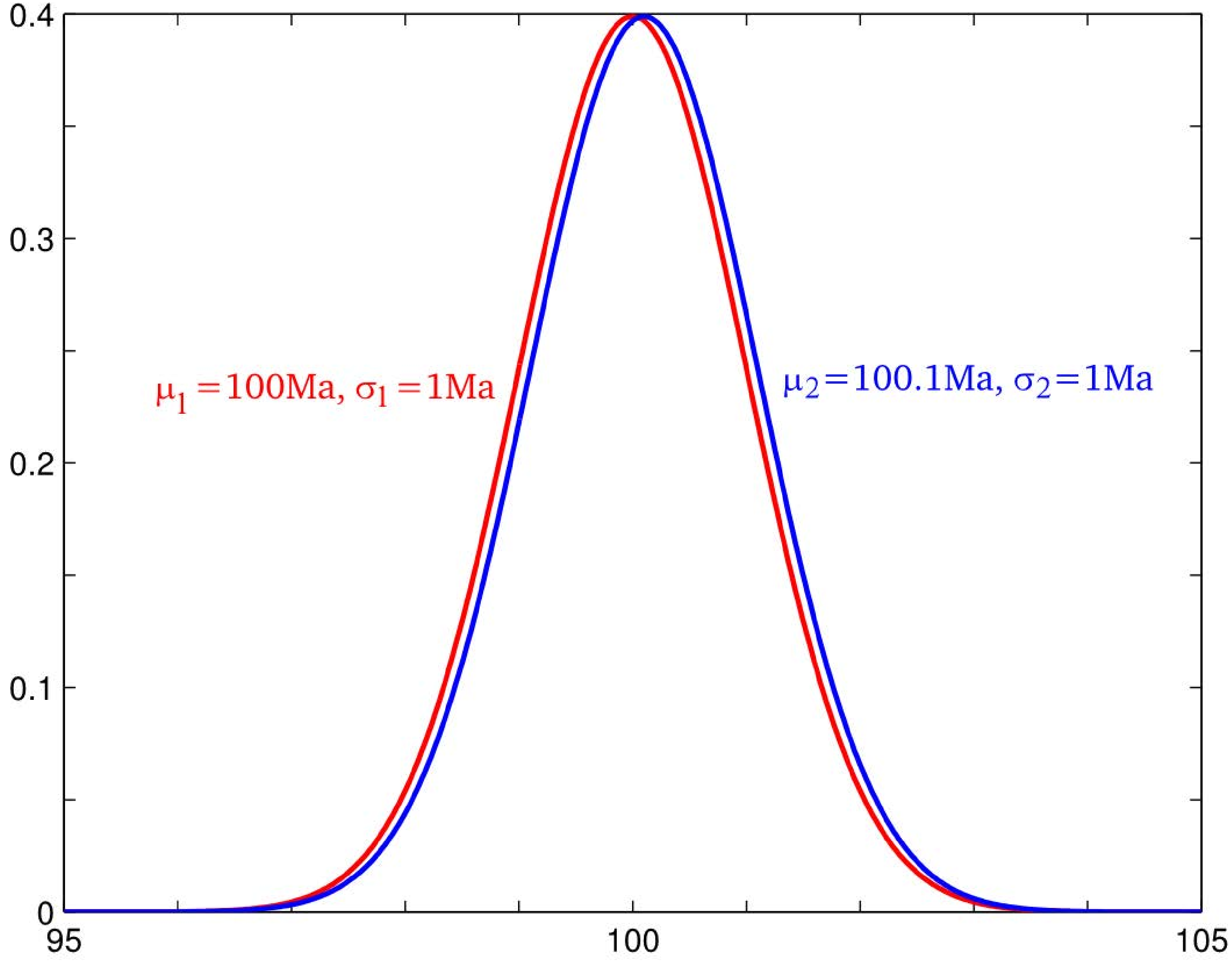
Any time
Since 2013
Since 2012
Since 2009
Custom range...

Sort by relevance

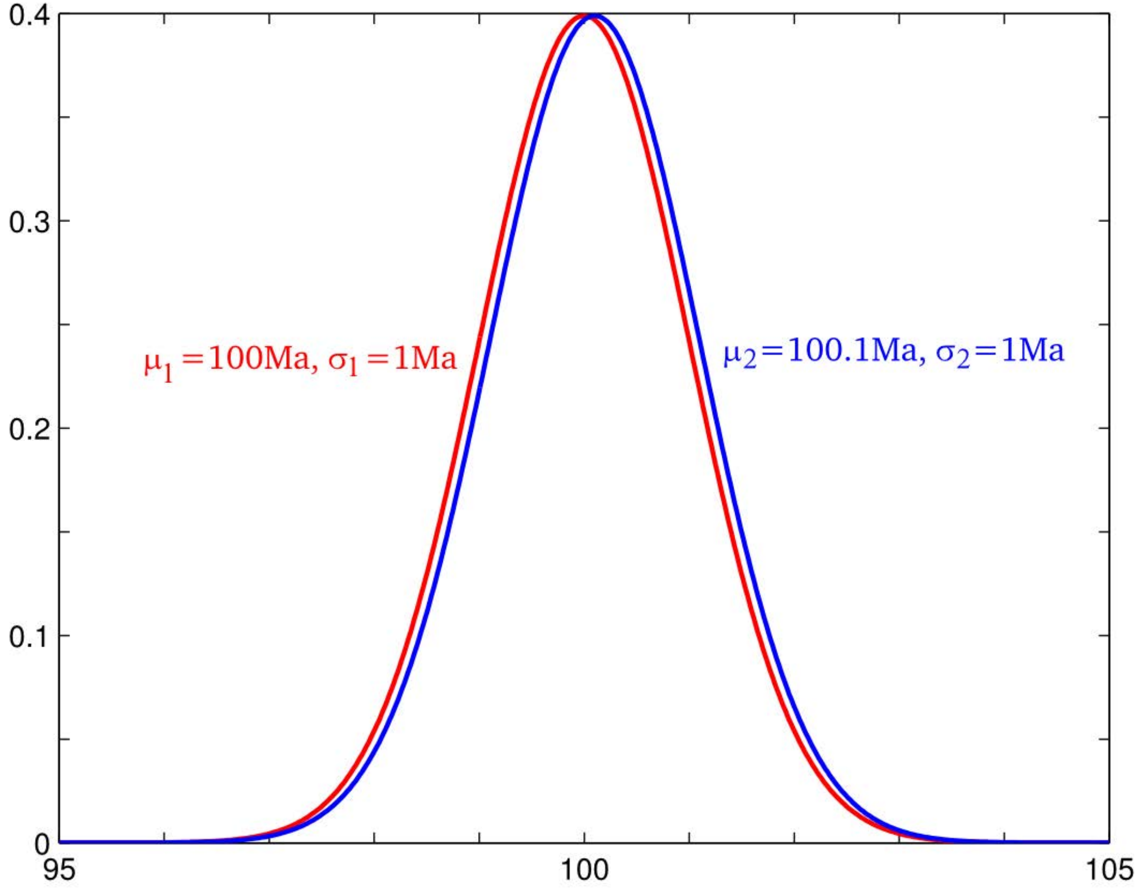
[BOOK] **Statistical power analysis** for the **behavioral sciences**
J **Cohen** - 1988 - books.google.com
Statistical Power Analysis is a nontechnical guide to **power analysis** in research planning that provides users of applied statistics with the tools they need for more effective **analysis**. The Second Edition includes:* a chapter covering **power analysis** in set correlation and **...**
Cited by 61850 Related articles All 16 versions Cite More

A **power** primer.
J **Cohen** - Psychological bulletin, 1992 - psycnet.apa.org
QUANTITATIVE METHODS IN PSYCHOLOGY A **Power** Primer **Jacob Cohen** New \brk University One possible reason for the continued neglect of **statistical power analysis** in research in the **behavioral sciences** is the inaccessibility of or difficulty with the standard material. **...**
Cited by 10626 Related articles All 60 versions Cite

An example: can you tell the difference between these two plutons?



Statistic: $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(S_1^2 + S_2^2)/N}}$ Effect size: $d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(S_1^2 + S_2^2)/2}}$ =0.1



N	t	p
10	$1/\sqrt{10}$	0.41
100	1	0.24
1,000	$\sqrt{10}$	0.013
10,000	100	$8\text{x}10^{-13}$

estimated bin proportions

$$\chi^2 = n\sum_i^k \frac{(\hat{p}_i^A - \hat{p}_i)^2}{\hat{p}_i} + m\sum_i^k \frac{(\hat{p}_i^B - \hat{p}_i)^2}{\hat{p}_i}$$

'average' bin proportions

bin proportions under H_a

$$\epsilon^2 = \sum_{i=1}^k \frac{(p_i^A - p_i)^2}{p_i} + \sum_{i=1}^k \frac{(p_i^B - p_i)^2}{p_i}$$

bin proportions under H_o

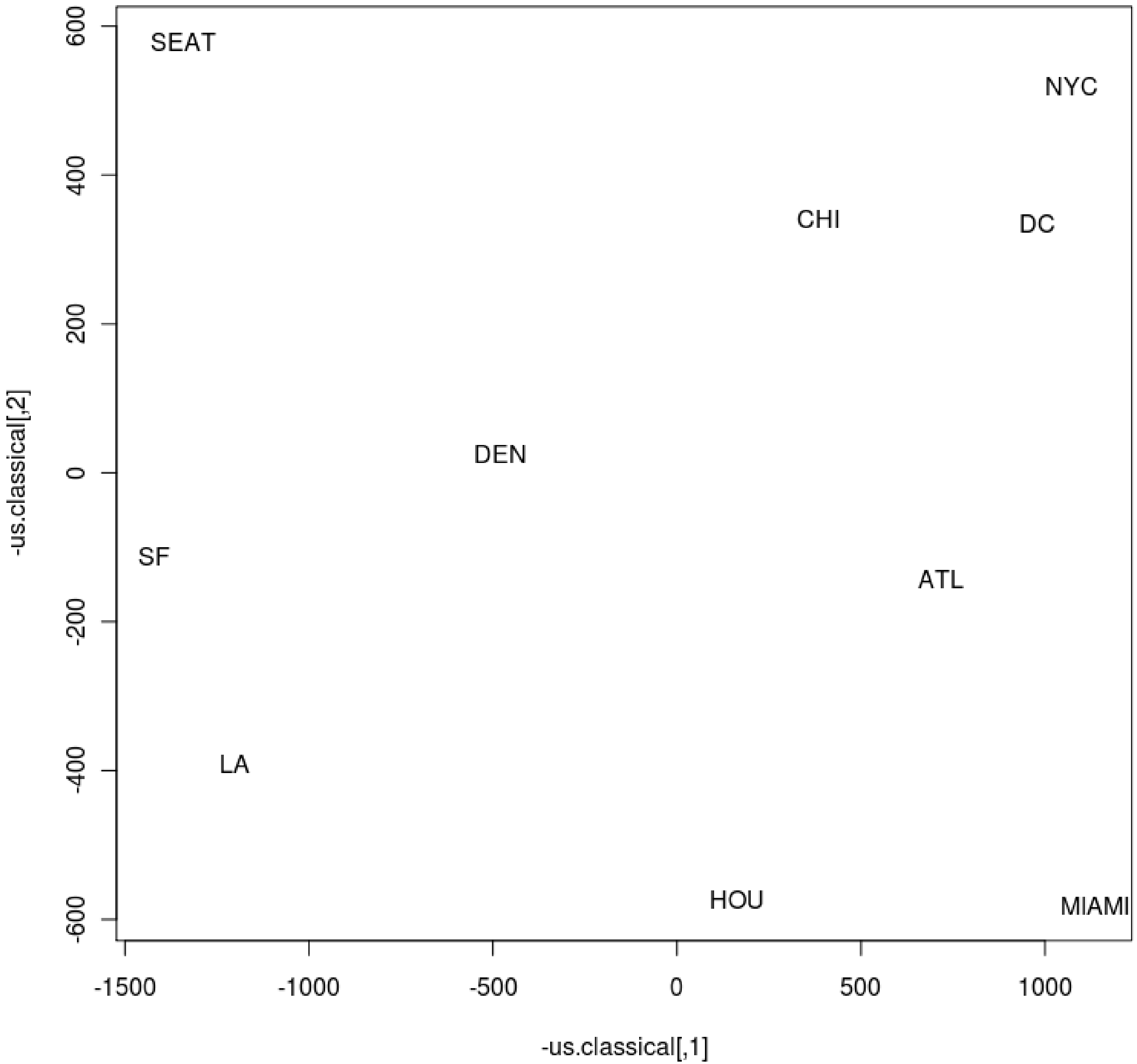
Conclusion: the best way to compare two age distributions is...

p-value → always scales with sample size

statistic → often scales with sample size

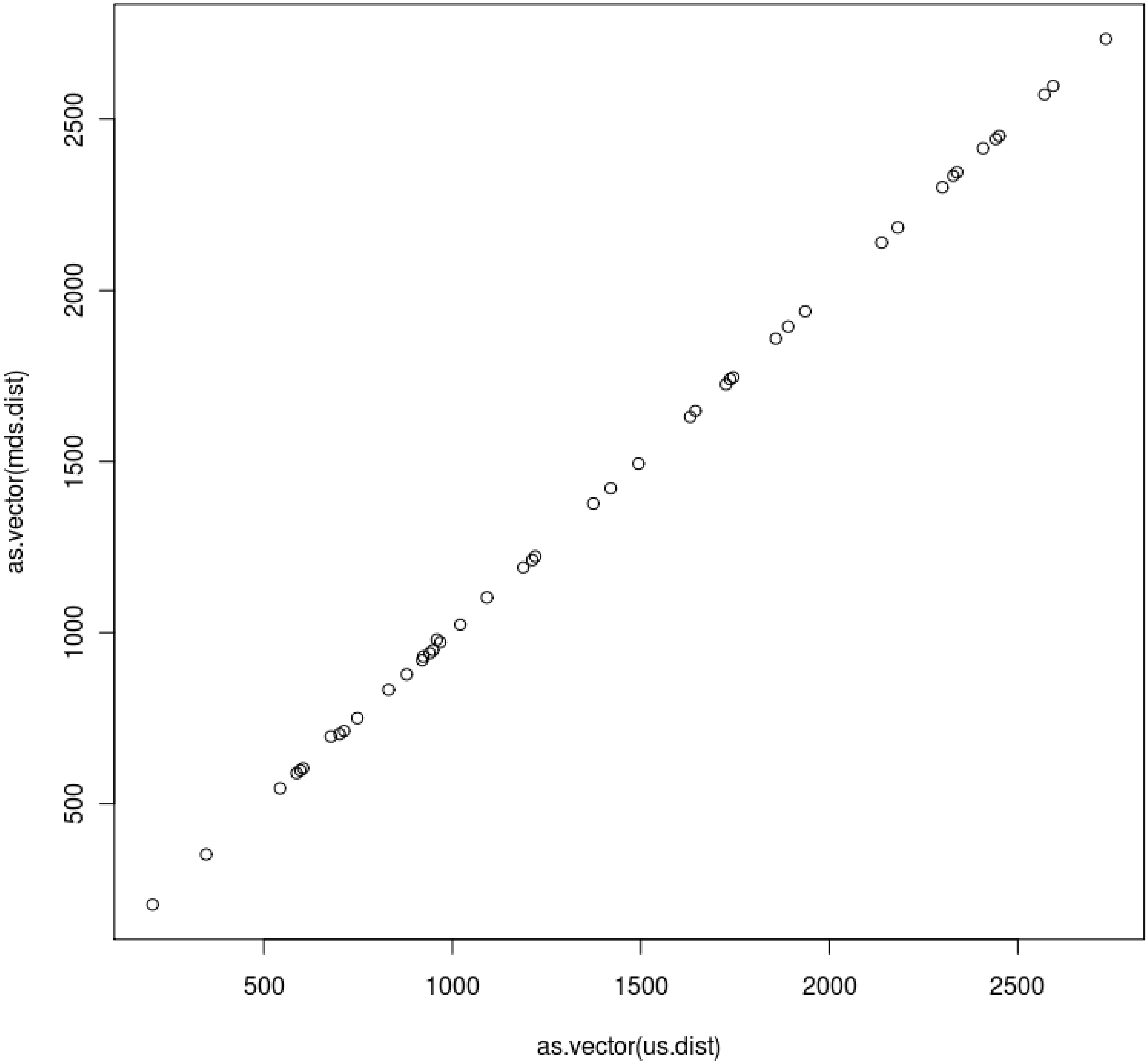
effect size → is independent of sample size

	ATL	CHI	DEN	HOU	LA	MIAMI	NYC	SF	SEAT	DC
ATL	0	587	1212	701	1936	604	748	2139	2182	543
CHI	587	0	920	940	1745	1188	713	1858	1737	597
DEN	1212	920	0	879	831	1726	1631	949	1021	1494
HOU	701	940	879	0	1374	968	1420	1645	1891	1220
LA	1936	1745	831	1374	0	2339	2451	347	959	2300
MIAMI	604	1188	1726	968	2339	0	1092	2594	2734	923
NYC	748	713	1631	1420	2451	1092	0	2571	2408	205
SF	2139	1858	949	1645	347	2594	2571	0	678	2442
SEAT	2182	1737	1021	1891	959	2734	2408	678	0	2329
DC	543	597	1494	1220	2300	923	205	2442	2329	0


SEAT
NYC
CHI
DC
DEN
SF
ATL
LA
HOU
MIAMI
600
400
200
0
-200
-400
-600
-us.classical[,2]
-1500
-1000
-500
0
500
1000
-us.classical[,1]

The “Shepard plot” is obtained by plotting dissimilarities vs. fitted distances

$$d_{i,j} = \sqrt{(x_i^1 - x_j^1)^2 + (x_i^2 - x_j^2)^2 + \ldots + (x_i^R - x_j^R)^2}$$



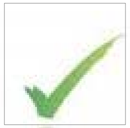
UCL

In order for the dissimilarity between samples i and j ($\delta_{i,j}$) to be a *metric*:

$$\delta_{i,j} = 0 \text{ if } i = j \text{ and } \delta_{i,j} > 0 \text{ otherwise} \qquad \text{(nonnegativity)}$$

$$\delta_{i,j} = \delta_{j,i} \qquad \text{(symmetry)}$$

$$\delta_{i,k} \leq \delta_{i,j} + \delta_{j,k} \qquad \text{(triangle inequality)}$$

 Chi-square

 Kolmogorov – Smirnov

 Cramér-von-Mises

Using a data transformation increases the robustness and reliability of MDS

disparity → $d_{i,j} \approx f(\delta_{i,j})$ ← dissimilarity

disparity transformation: (arrow to f)

1. $d_{i,j} = \delta_{i,j}$ — classical MDS

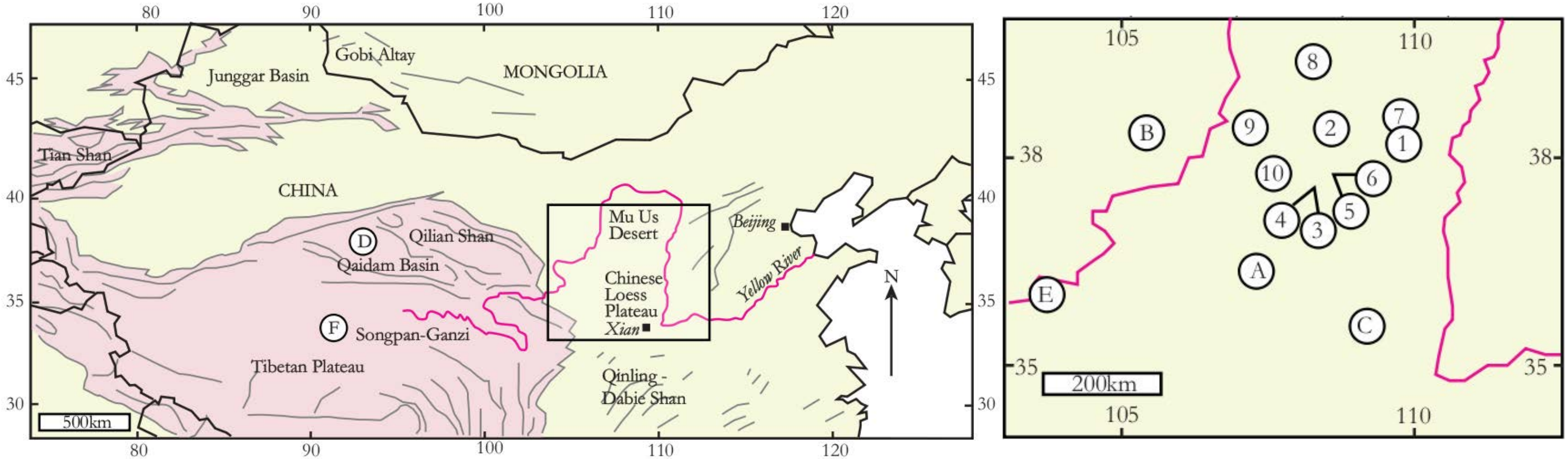
2. $d_{i,j} = a\delta_{i,j}$
   $d_{i,j} = a + b\delta_{i,j}$ — metric MDS
   $d_{i,j} = e^{\delta_{i,j}}$
   …

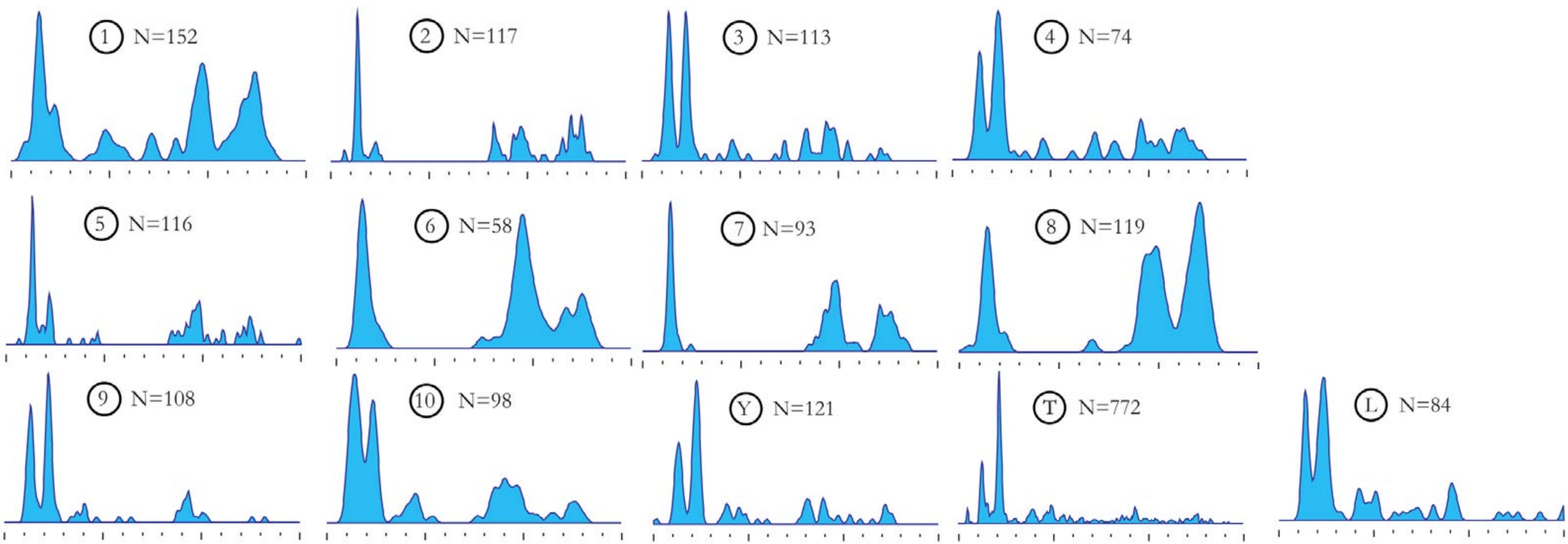
3. $f(d_{i,j}) > f(d_{k,l})$ if $\delta_{i,j} > \delta_{k,l}$ — nonmetric MDS

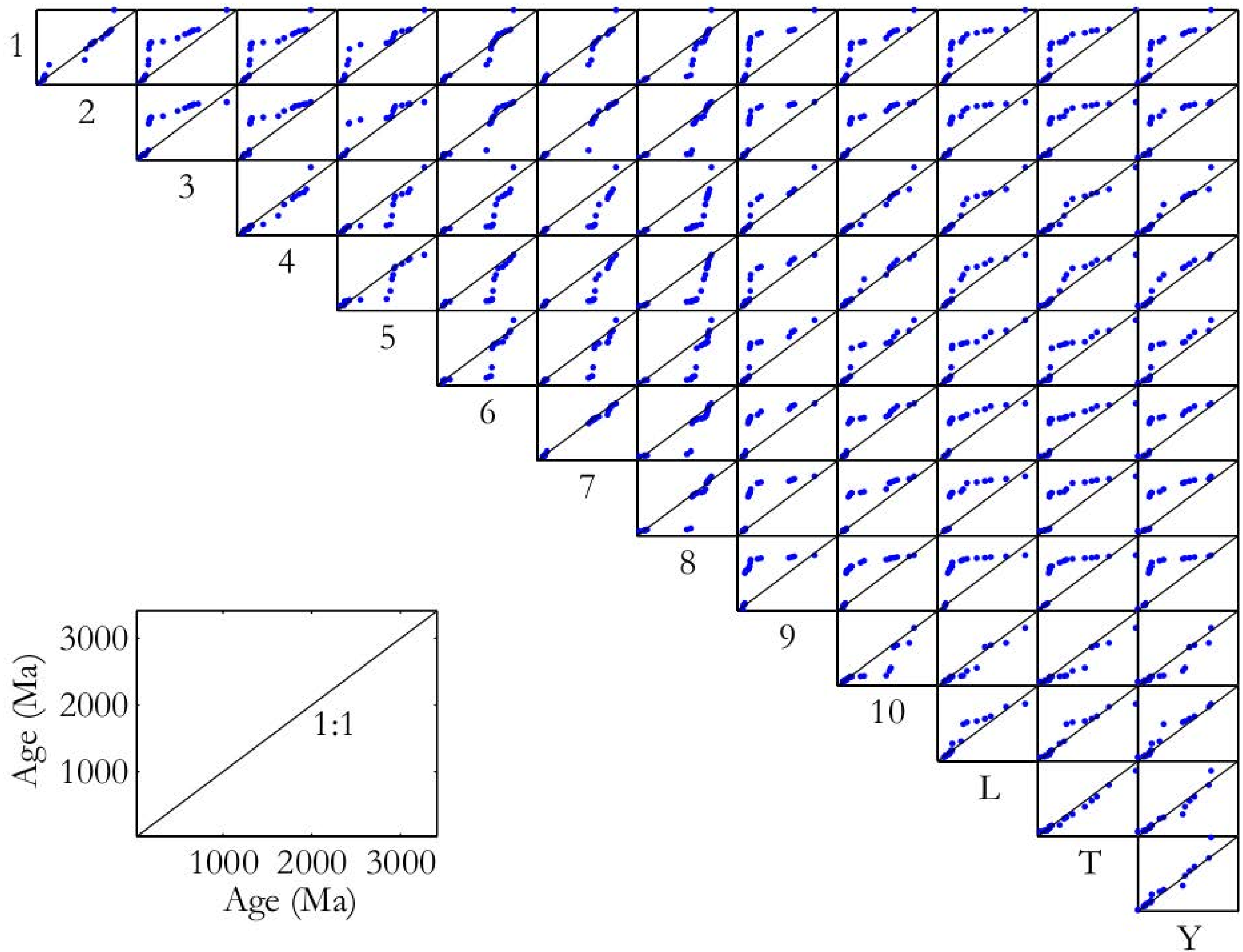
Classical MDS Metric MDS Nonmetric MDS

$$S = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}[f(\delta_{i,j}) - d_{i,j}]^2}{\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{i,j}^2}}$$

g.o.f	poor	fair	good	excellent	perfect
S	0.2	0.1	0.05	0.025	0


Gobi Altay
Junggar Basin
MONGOLIA
Tian Shan
CHINA
Mu Us Desert
Beijing
Qilian Shan
Qaidam Basin
Chinese Loess Plateau
Yellow River
N
Xian
Songpan-Ganzi
Tibetan Plateau
Qinling - Dabie Shan
500km
200km


1 N=152
2 N=117
3 N=113
4 N=74
5 N=116
6 N=58
7 N=93
8 N=119
9 N=108
10 N=98
Y N=121
T N=772
L N=84


1
2
3
4
5
6
7
8
9
10
L
T
Y
3000
2000
1000
Age (Ma)
1:1
1000 2000 3000
Age (Ma)

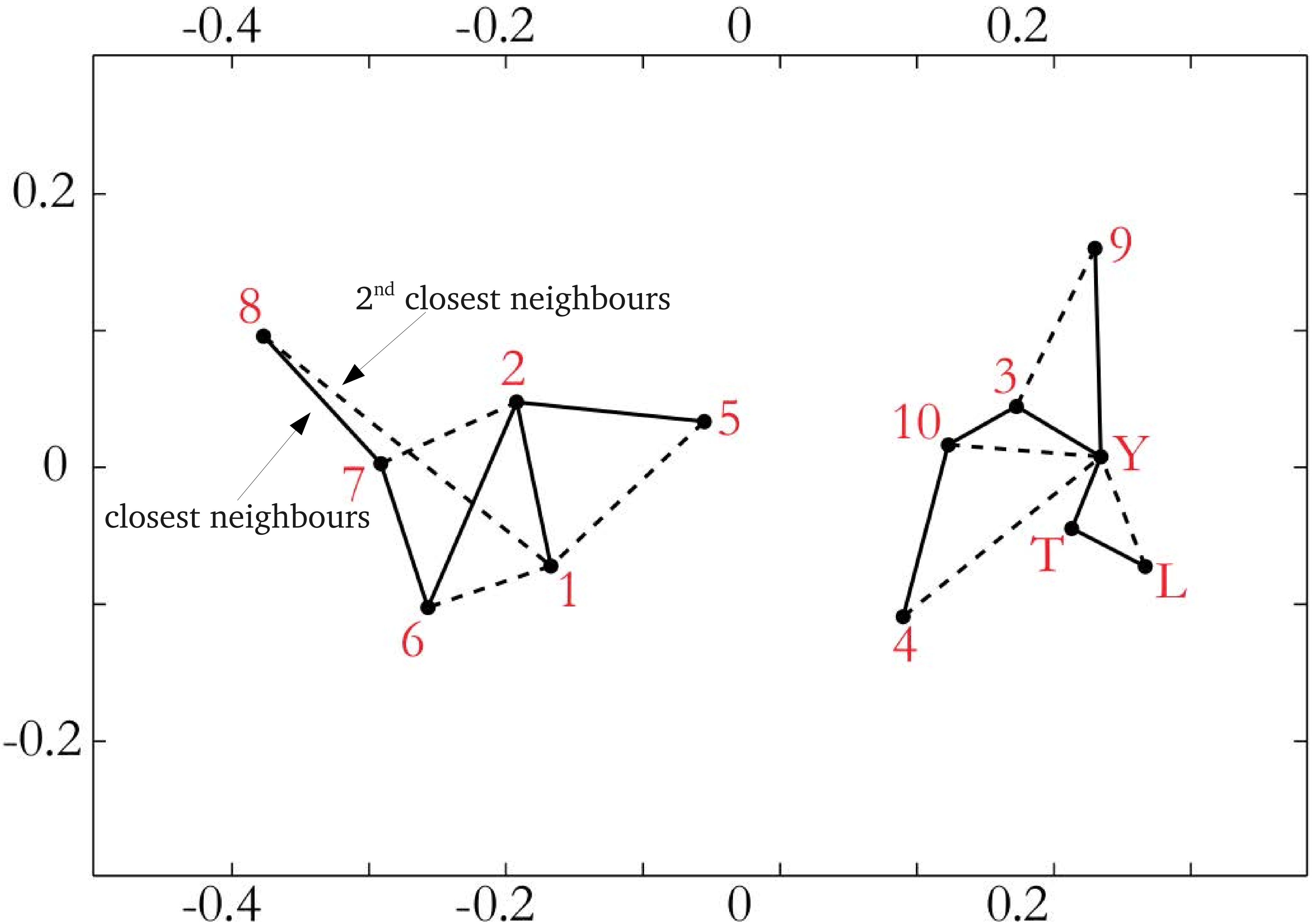
$\delta =$

	1	2	3	4	5	6	7	8	9	10	L	T	Y
1	0	14	33	27	18	14	15	22	48	32	42	37	40
2	14	0	36	33	16	14	15	24	46	32	47	42	43
3	33	36	0	19	24	44	47	55	17	10	13	12	8
4	27	33	19	0	20	38	41	48	28	14	21	17	16
5	18	16	24	20	0	22	24	33	31	20	33	28	30
6	14	14	44	38	22	0	14	24	52	41	52	48	49
7	15	15	47	41	24	14	0	16	51	43	54	49	52
8	22	24	55	48	33	24	16	0	61	53	63	59	62
9	48	46	17	28	31	52	51	61	0	20	22	18	16
10	32	32	10	14	20	41	43	53	20	0	17	15	13
L	42	47	13	21	33	52	54	63	22	17	0	10	11
T	37	42	12	17	28	48	49	59	18	15	10	0	7
Y	40	43	8	16	30	49	52	62	16	13	11	7	0

8 dimensions are enough to completely describe the data (i.e., S=0)

$\mathbf{x} =$

	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
1	−17	10.0	−1.9	2.6	−2.2	1.1	1.2	0.63
2	−19	−2.0	7.4	−2.4	−5.2	0.1	0.8	−0.39
3	17	0.1	2.2	2.8	−6.9	2.0	−1.6	0.19
4	9	7.0	−4.0	−8.4	3.2	0.1	0.9	−0.93
5	−5	−3.8	1.7	−2.6	2.8	−1.2	−2.9	2.58
6	−25	2.1	8.2	2.6	6.4	1.3	−2.6	−0.45
7	−28	−4.7	−1.6	4.1	2.1	−2.3	4.0	−1.51
8	−37	−2.7	−10.3	−2.2	−3.0	0.9	−1.9	0.13
9	23	−16.2	−2.6	0.3	1.3	1.0	0.2	−0.31
10	14	0.1	5.9	−3.3	−2.0	−4.1	2.0	1.03
L	25	5.1	−4.1	4.0	−0.1	−4.6	−3.1	−1.44
T	21	3.4	−3.3	3.6	2.6	2.0	2.6	2.84
Y	23	1.6	2.3	−1.1	0.9	3.7	0.3	−2.37

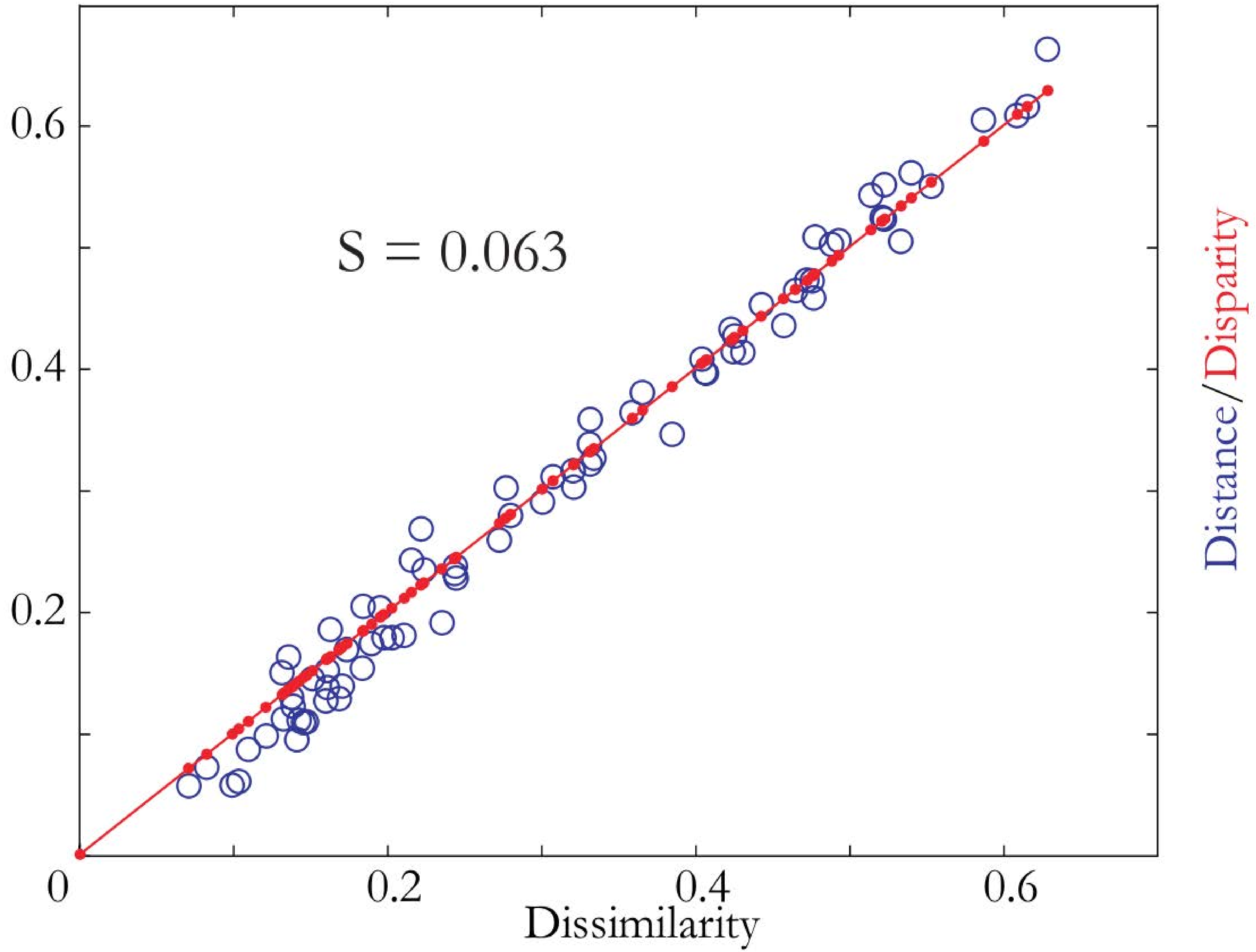

-0.4
-0.2
0
0.2
0.2
0
-0.2
8
2nd closest neighbours
2
5
7
closest neighbours
1
6
9
3
10
Y
T
L
4

The fitted distances are in good agreement with the dissimilarities from 3 slides ago

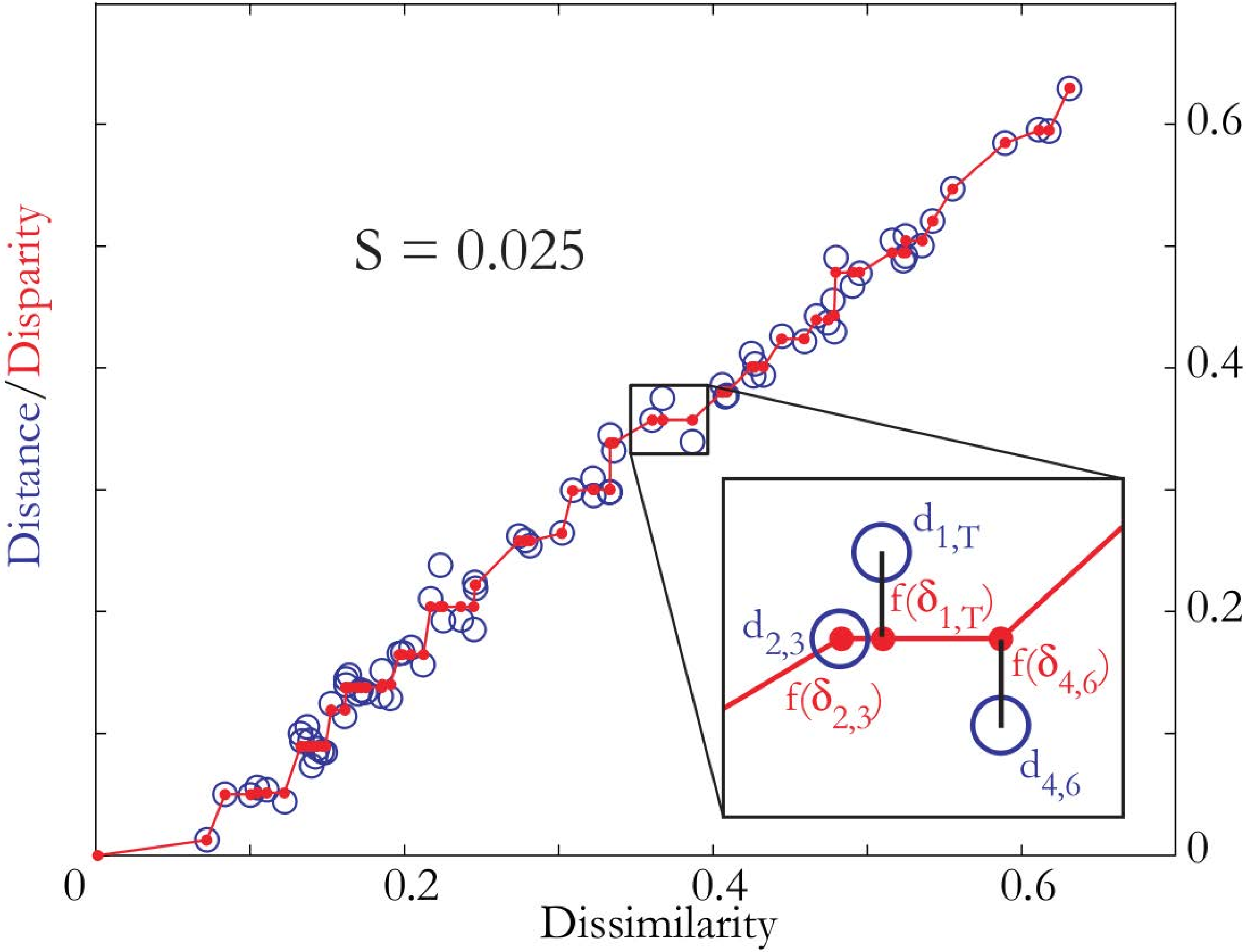
$d =$

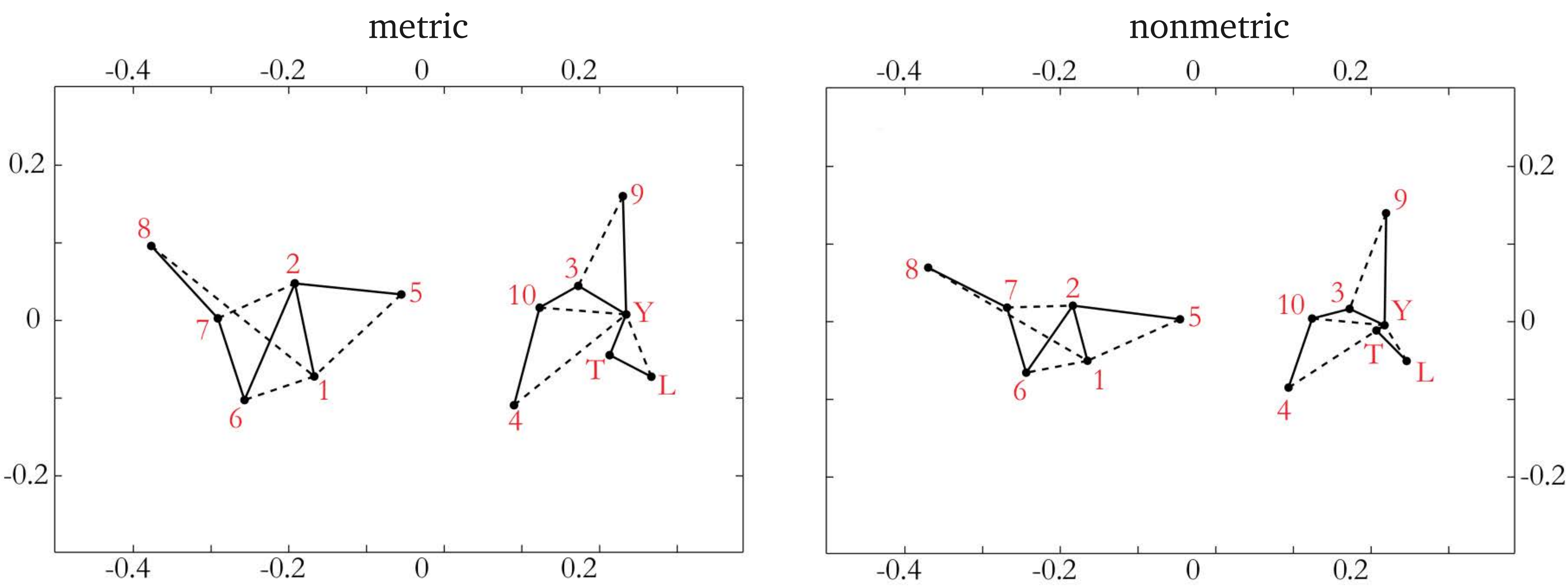
	1	2	3	4	5	6	7	8	9	10	L	T	Y
1	0	12	35	26	18	11	18	23	48	32	42	38	41
2	12	0	36	30	14	7	9	18	44	33	45	40	43
3	35	36	0	10	23	42	45	54	17	3	9	5	7
4	26	30	10	0	18	35	39	47	27	8	16	12	15
5	18	14	23	18	0	21	22	31	31	20	32	27	29
6	11	7	42	35	21	0	7	13	51	39	50	46	48
7	18	9	45	39	22	7	0	9	52	42	54	49	51
8	23	18	54	47	31	13	9	0	61	50	62	58	60
9	48	44	17	27	31	51	52	61	0	19	21	20	18
10	32	33	3	8	20	39	42	50	19	0	12	8	10
L	42	45	9	16	32	50	54	62	21	12	0	5	4
T	38	40	5	12	27	46	49	58	20	8	5	0	3
Y	41	43	7	15	29	48	51	60	18	10	4	3	0

The Shepard plot and stress value indicate a 'good' fit for the metric MDS



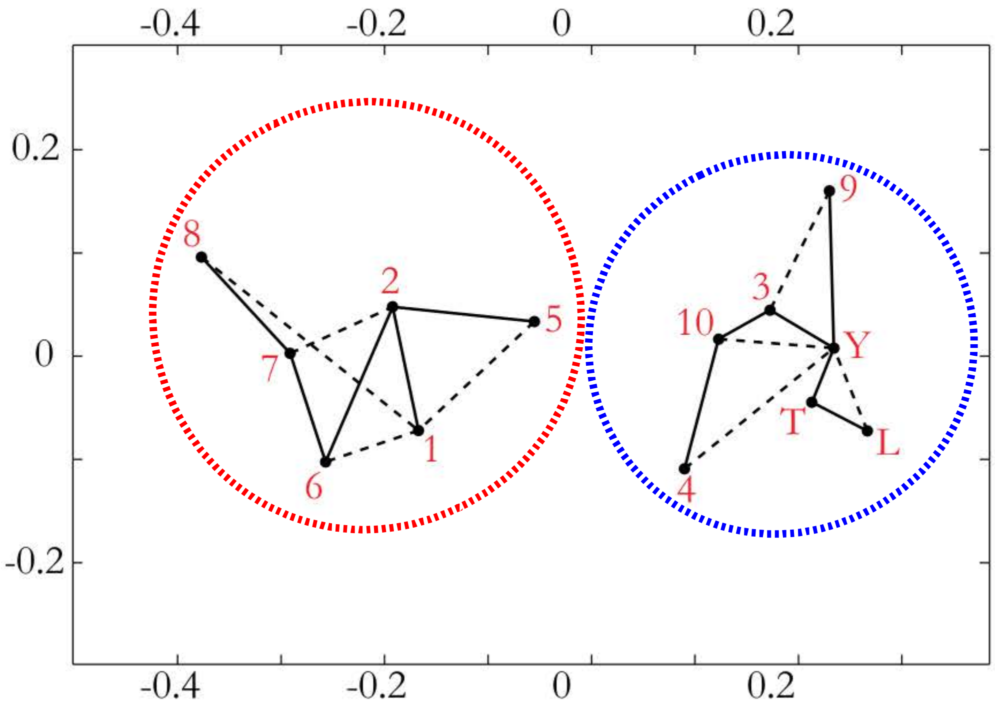
The fit and stress value of the nonmetric MDS, however, qualify as 'excellent'



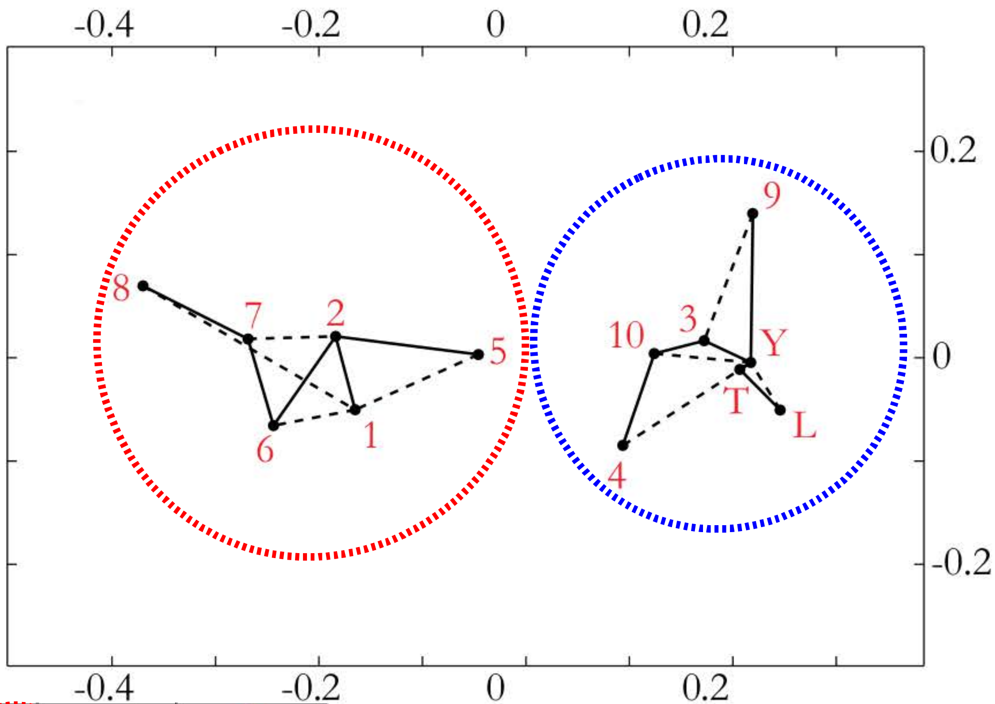

metric
nonmetric
-0.4
-0.2
0
0.2
-0.2
8
7
2
5
6
1
10
3
9
Y
T
L
4

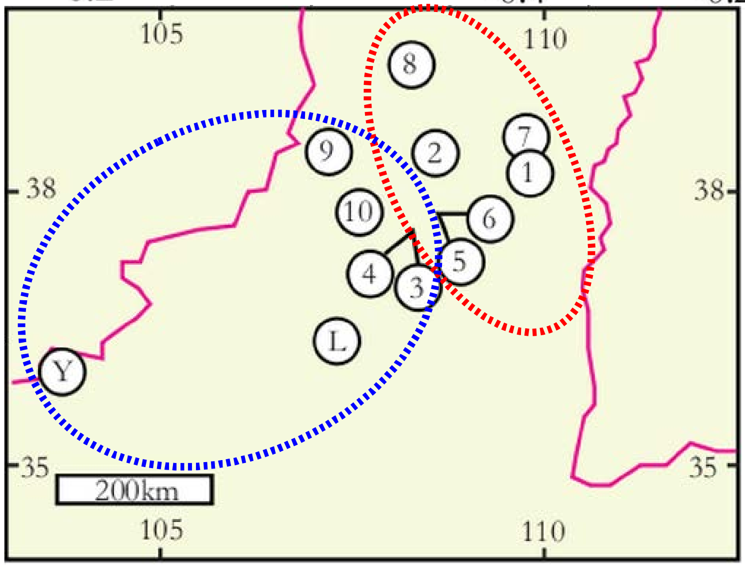
MDS analysis reveals a pronounced East-West dichotomy of the field area

metric



nonmetric

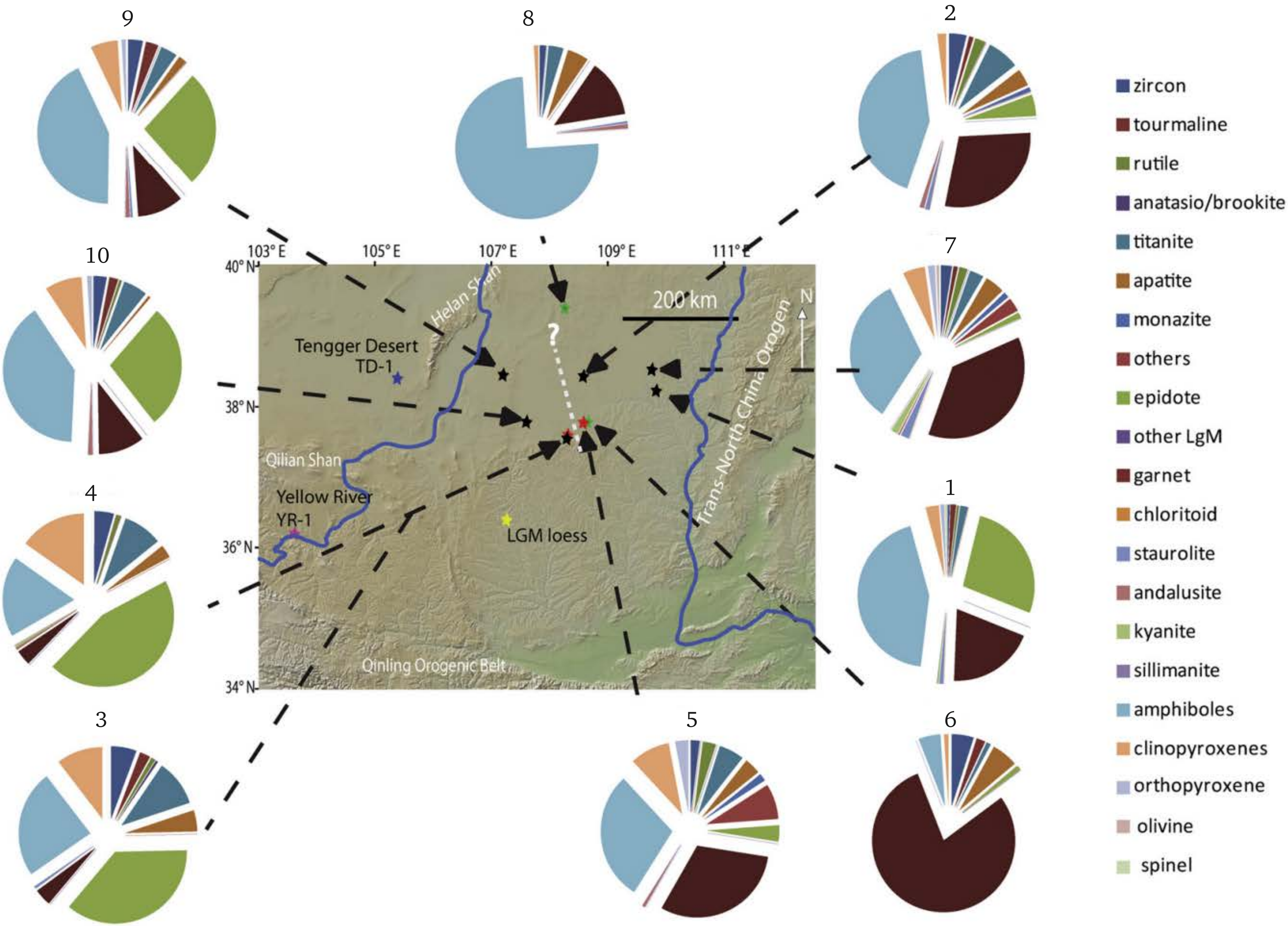




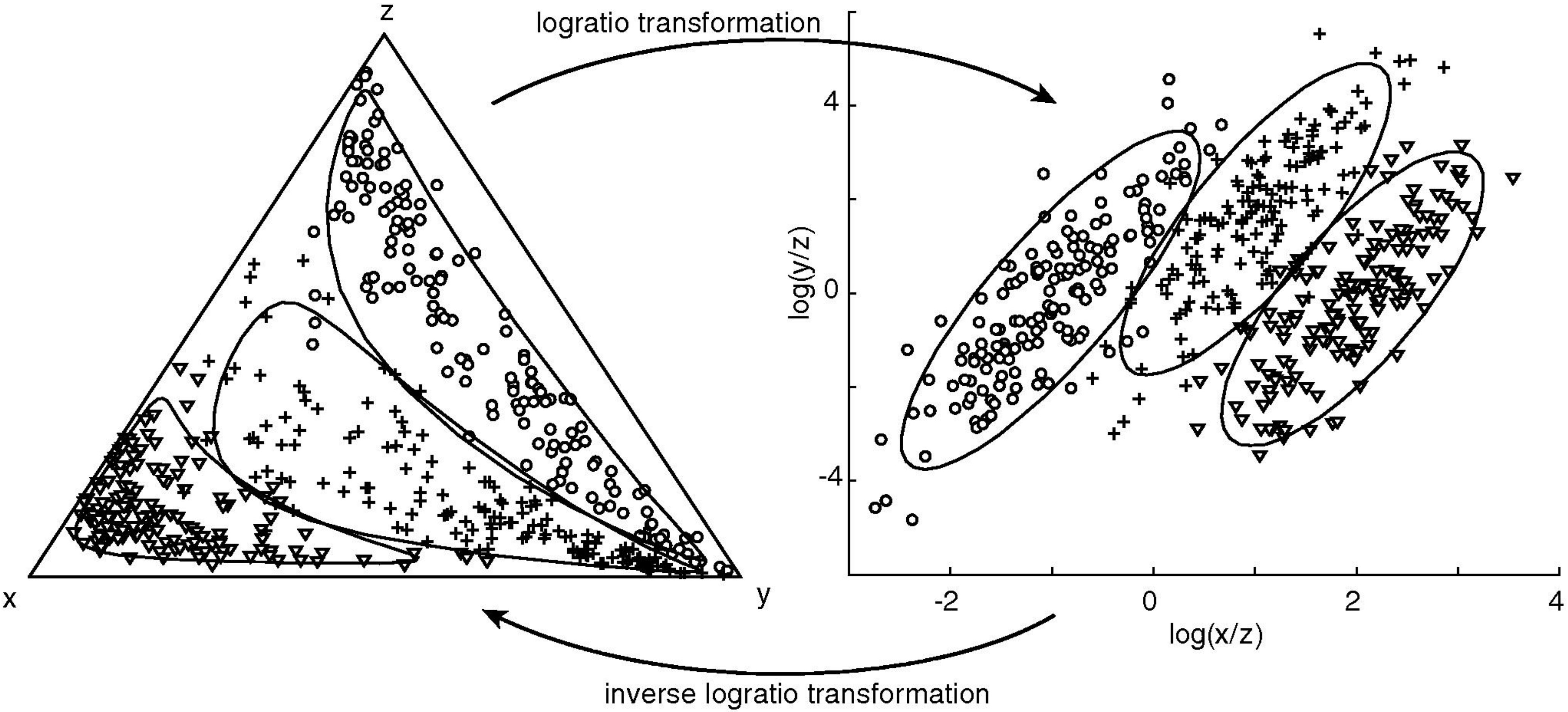
PCA = classical MDS with

$$d_{i,j} = \sqrt{(x_i^1 - x_j^1)^2 + (x_i^2 - x_j^2)^2 + \ldots + (x_i^R - x_j^R)^2}$$

PCA Classical MDS Metric MDS Nonmetric MDS


9
8
2
10
7
4
1
3
5
6
103° E
105° E
107° E
109° E
111° E
40° N
38° N
36° N
34° N
Helan Shan
Tengger Desert
TD-1
200 km
N
Trans-North China Orogen
Qilian Shan
Yellow River
YR-1
LGM loess
Qinling Orogenic Belt
?
zircon
tourmaline
rutile
anatasio/brookite
titanite
apatite
monazite
others
epidote
other LgM
garnet
chloritoid
staurolite
andalusite
kyanite
sillimanite
amphiboles
clinopyroxenes
orthopyroxene
olivine
spinel

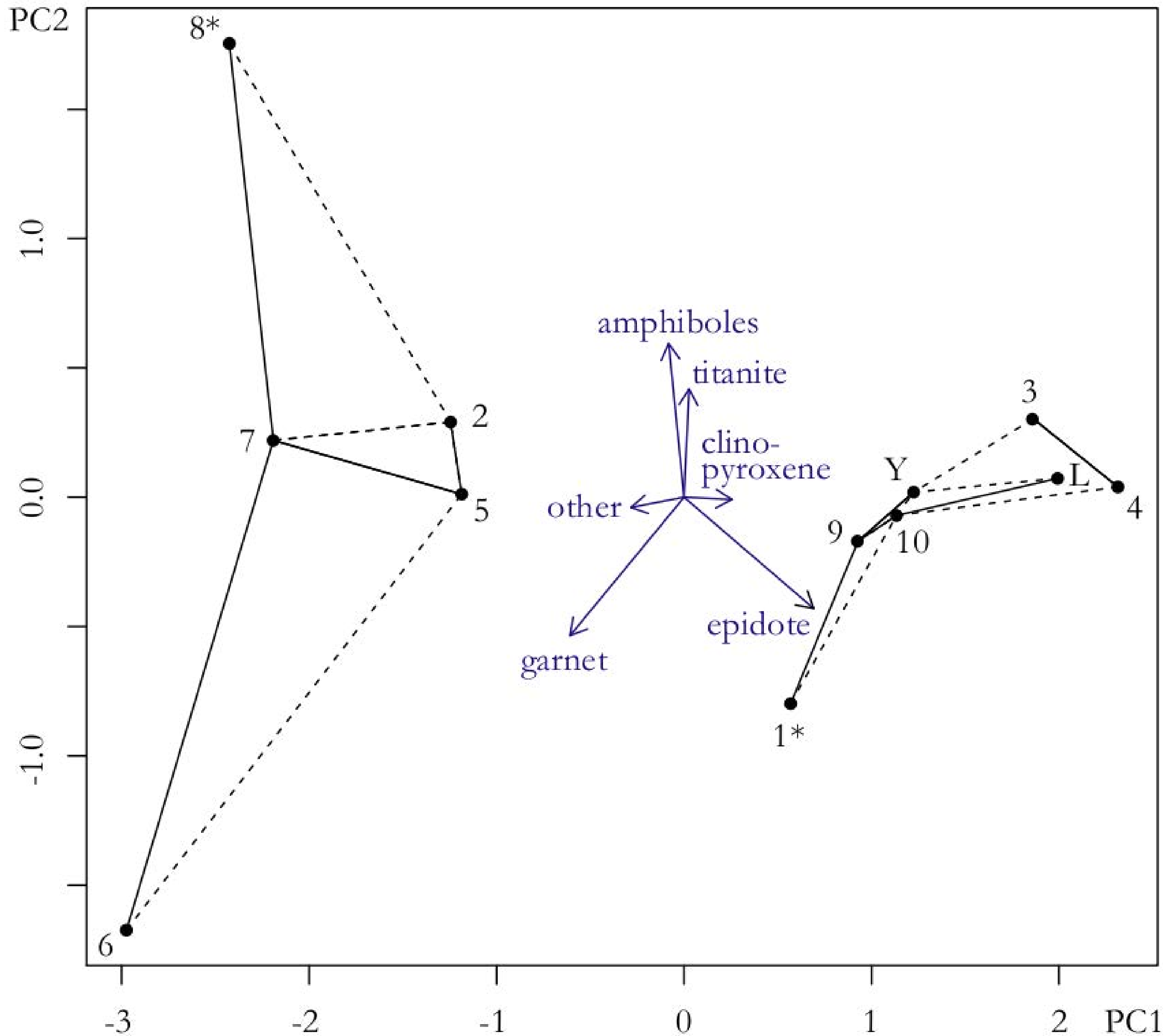
		sph	ep	gt	am	cpx	other
	1	4	64	46	103	7	11
	2	15	10	60	89	4	29
	3	22	77	8	51	22	31
	4	18	92	7	37	31	19
i	5	17	11	93	88	27	66
	6	2	2	160	10	2	26
	7	10	4	119	108	16	64
j	8	7	1	27	157	2	16
	9	8	57	22	91	13	22
	10	11	57	21	81	17	16
	Y	17	124	29	176	26	47
	L	7	67	11	85	29	7


z
x
y
logratio transformation
inverse logratio transformation
4
0
-4
log(y/z)
-2
0
2
4
log(x/z)

		sph	ep	gt	am	cpx	other
	1	4	64	46	103	7	11
	2	15	10	60	89	4	29
	3	22	77	8	51	22	31
	4	18	92	7	37	31	19
i	5	17	11	93	88	27	66
	6	2	2	160	10	2	26
	7	10	4	119	108	16	64
j	8	7	1	27	157	2	16
	9	8	57	22	91	13	22
	10	11	57	21	81	17	16
	Y	17	124	29	176	26	47
	L	7	67	11	85	29	7

$$\delta_{i,j} = \sqrt{\ln\left(\frac{sph_i}{sph_j}\right)^2 + \ldots + \ln\left(\frac{oth_i}{oth_j}\right)^2}$$

		1	2	3	4	5	6	7	8 (j)	9	10	Y	L
	1	0	2.56	2.91	3.06	2.97	4.06	3.57	4.03	1.32	1.56	1.67	2.15
	2	2.56	0	3.34	3.81	1.62	2.76	1.89	2.04	2.39	2.54	2.65	3.61
	3	2.91	3.34	0	0.70	3.20	5.26	4.22	4.76	1.68	1.48	1.36	1.88
	4	3.06	3.81	0.70	0	3.55	5.59	4.60	5.25	1.95	1.64	1.67	1.65
i	5	2.97	1.62	3.20	3.55	0	2.57	1.11	2.57	2.49	2.55	2.72	3.42
	6	4.06	2.76	5.26	5.59	2.57	0	2.22	3.58	4.25	4.45	4.58	5.41
	7	3.57	1.89	4.22	4.60	1.11	2.22	0	2.00	3.33	3.47	3.59	4.30
	8	4.03	2.04	4.76	5.25	2.57	3.58	2.00	0	3.88	4.02	4.08	4.82
	9	1.32	2.39	1.68	1.95	2.49	4.25	3.33	3.88	0	0.54	0.42	1.51
	10	1.56	2.54	1.48	1.64	2.55	4.45	3.47	4.02	0.54	0	0.65	1.18
	Y	1.67	2.65	1.36	1.67	2.72	4.58	3.59	4.08	0.42	0.65	0	1.45
	L	2.15	3.61	1.88	1.65	3.42	5.41	4.30	4.82	1.51	1.18	1.45	0



PC2
8*
1.0
amphiboles
titanite
clino-
pyroxene
3
2
7
Y
L
0.0
other
5
4
9
10
epidote
garnet
1*
-1.0
6
-3
-2
-1
0
1
2
PC1

Multi-Dimensional Scali… ×

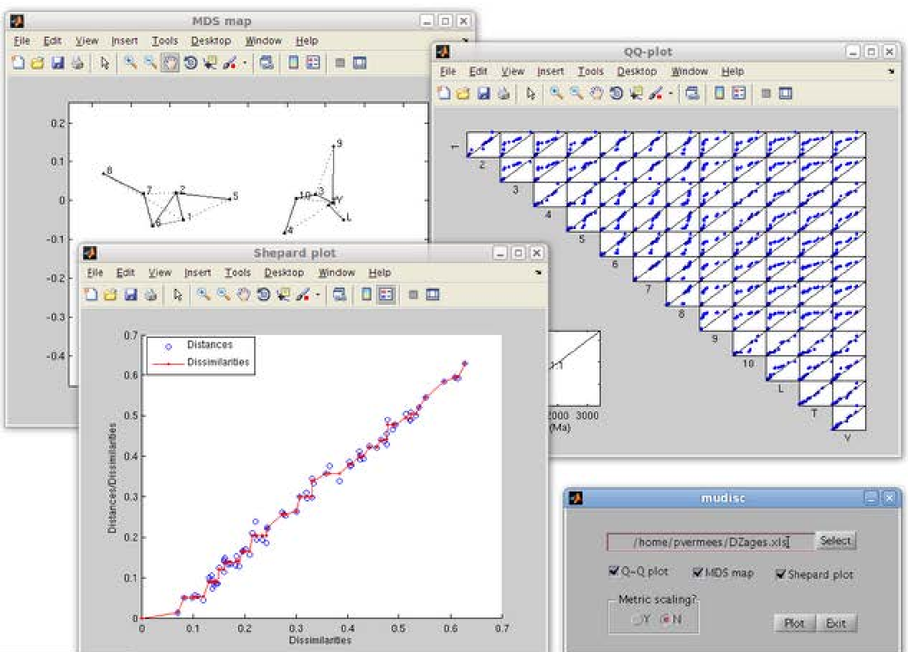
www.ucl.ac.uk/~ucfbpve/mudisc/

MuDiSc: Multi-Dimensional Scaling with Matlab and R

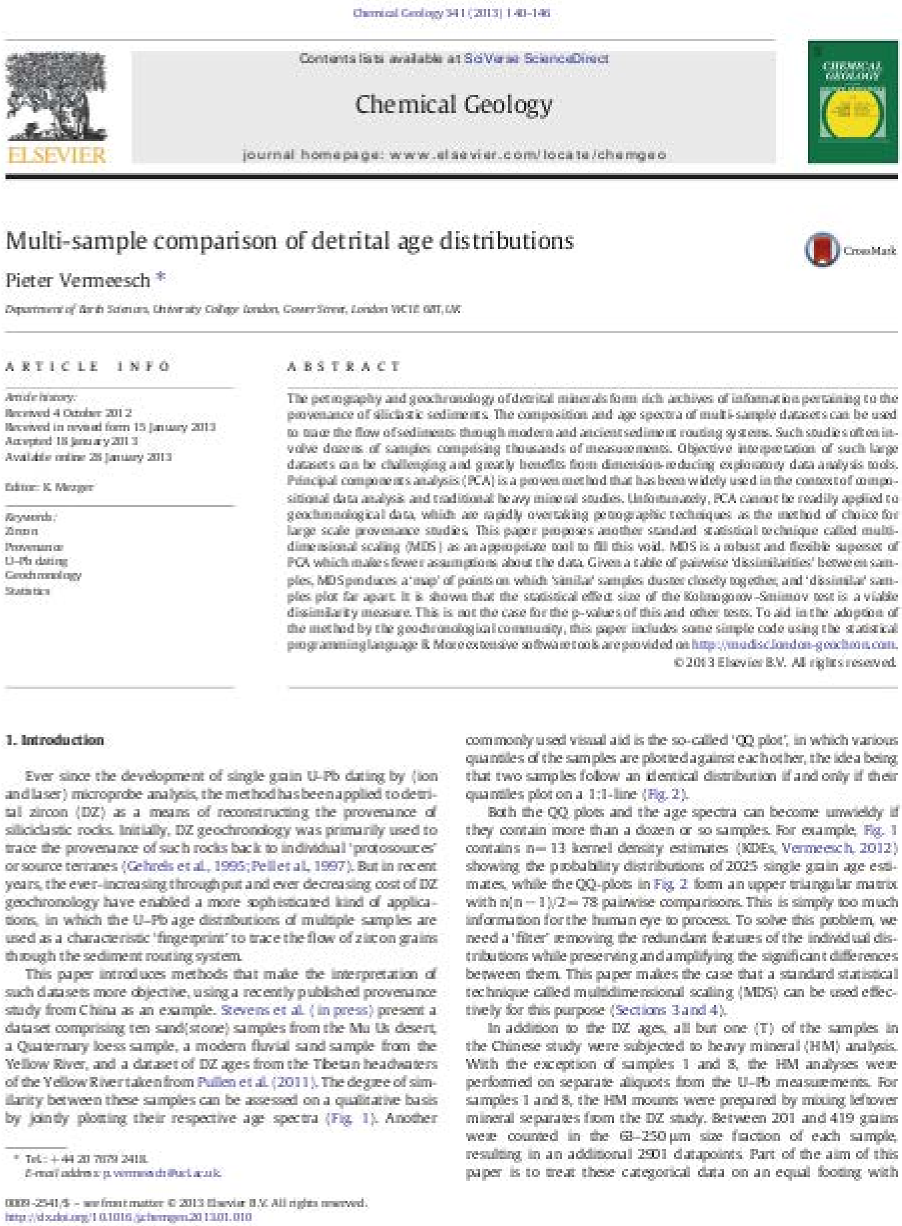
An increasing number of detrital zircon provenance studies are based on not just a few but many samples. This trend is likely to continue as the price of zircon U-Pb analyses continues to drop. The large datasets resulting from such studies call for a dimension-reducing technique such as Multi-Dimensional Scaling (MDS). Given a dissimilarity matrix (i.e., a table of pairwise distances), MDS constructs a 'map' on which 'similar' samples cluster closely together and 'dissimilar' samples plot far apart. This website presents some software tools for MDS analysis in the context of detrital geochronology, using the *effect size* of the two-sample Kolmogorov-Smirnov statistic as a dissimilarity metric. Two alternative sets of tools are presented here, written in `Matlab` (Section 1) and `R` (Section 2). MDS is closely related to, and in fact is a superset of, Principal Component Analysis (PCA), which is an established technique for conventional petrographic provenance studies. An example of this (written in `R`) is given at the end of this page, in Section 3. Further detail about these methods is provided in an accompanying paper:

Vermeesch, P., 2013, Multi-sample comparison of detrital age distributions. Chemical Geology, doi:10.1016/j.chemgeo.2013.01.010.

1. A user-friendly Matlab-GUI:



www.ucl.ac.uk/~ucfbpve/pictures/MuDiSc.png



Chemical Geology 341 (2013) 140–146

Contents lists available at SciVerse ScienceDirect

Chemical Geology

journal homepage: www.elsevier.com/locate/chemgeo

Multi-sample comparison of detrital age distributions

Pieter Vermeesch *

Department of Earth Sciences, University College London, Gower Street, London WC1E 6BT, UK

ARTICLE INFO

Article history:
Received 4 October 2012
Received in revised form 15 January 2013
Accepted 18 January 2013
Available online 28 January 2013

Editor: K. Mezger

Keywords:
Zircon
Provenance
U–Pb dating
Geochronology
Statistics

ABSTRACT

The petrography and geochronology of detrital minerals form rich archives of information pertaining to the provenance of siliclastic sediments. The composition and age spectra of multi-sample datasets can be used to trace the flow of sediments through modern and ancient sediment routing systems. Such studies often involve dozens of samples comprising thousands of measurements. Objective interpretation of such large datasets can be challenging and greatly benefits from dimension-reducing exploratory data analysis tools. Principal components analysis (PCA) is a proven method that has been widely used in the context of compositional data analysis and traditional heavy mineral studies. Unfortunately, PCA cannot be readily applied to geochronological data, which are rapidly overtaking petrographic techniques as the method of choice for large scale provenance studies. This paper proposes another standard statistical technique called multidimensional scaling (MDS) as an appropriate tool to fill this void. MDS is a robust and flexible superset of PCA which makes fewer assumptions about the data. Given a table of pairwise 'dissimilarities' between samples, MDS produces a 'map' of points on which 'similar' samples cluster closely together, and 'dissimilar' samples plot far apart. It is shown that the statistical effect size of the Kolmogorov–Smirnov test is a viable dissimilarity measure. This is not the case for the p-values of this and other tests. To aid in the adoption of the method by the geochronological community, this paper includes some simple code using the statistical programming language R. More extensive software tools are provided on http://mudisc.london-geochron.com.

1. Introduction

Ever since the development of single grain U–Pb dating by (ion and laser) microprobe analysis, the method has been applied to detrital zircon (DZ) as a means of reconstructing the provenance of siliciclastic rocks. Initially, DZ geochronology was primarily used to trace the provenance of such rocks back to individual 'protosources' or source terranes (Gehrels et al., 1995; Pell et al., 1997). But in recent years, the ever-increasing throughput and ever decreasing cost of DZ geochronology have enabled a more sophisticated kind of applications, in which the U–Pb age distributions of multiple samples are used as a characteristic 'fingerprint' to trace the flow of zircon grains through the sediment routing system.

This paper introduces methods that make the interpretation of such datasets more objective, using a recently published provenance study from China as an example. Stevens et al. (in press) present a dataset comprising ten sand(stone) samples from the Mu Us desert, a Quaternary loess sample, a modern fluvial sand sample from the Yellow River, and a dataset of DZ ages from the Tibetan headwaters of the Yellow River taken from Pullen et al. (2011). The degree of similarity between these samples can be assessed on a qualitative basis by jointly plotting their respective age spectra (Fig. 1). Another commonly used visual aid is the so-called 'QQ plot', in which various quantiles of the samples are plotted against each other, the idea being that two samples follow an identical distribution if and only if their quantiles plot on a 1:1-line (Fig. 2).

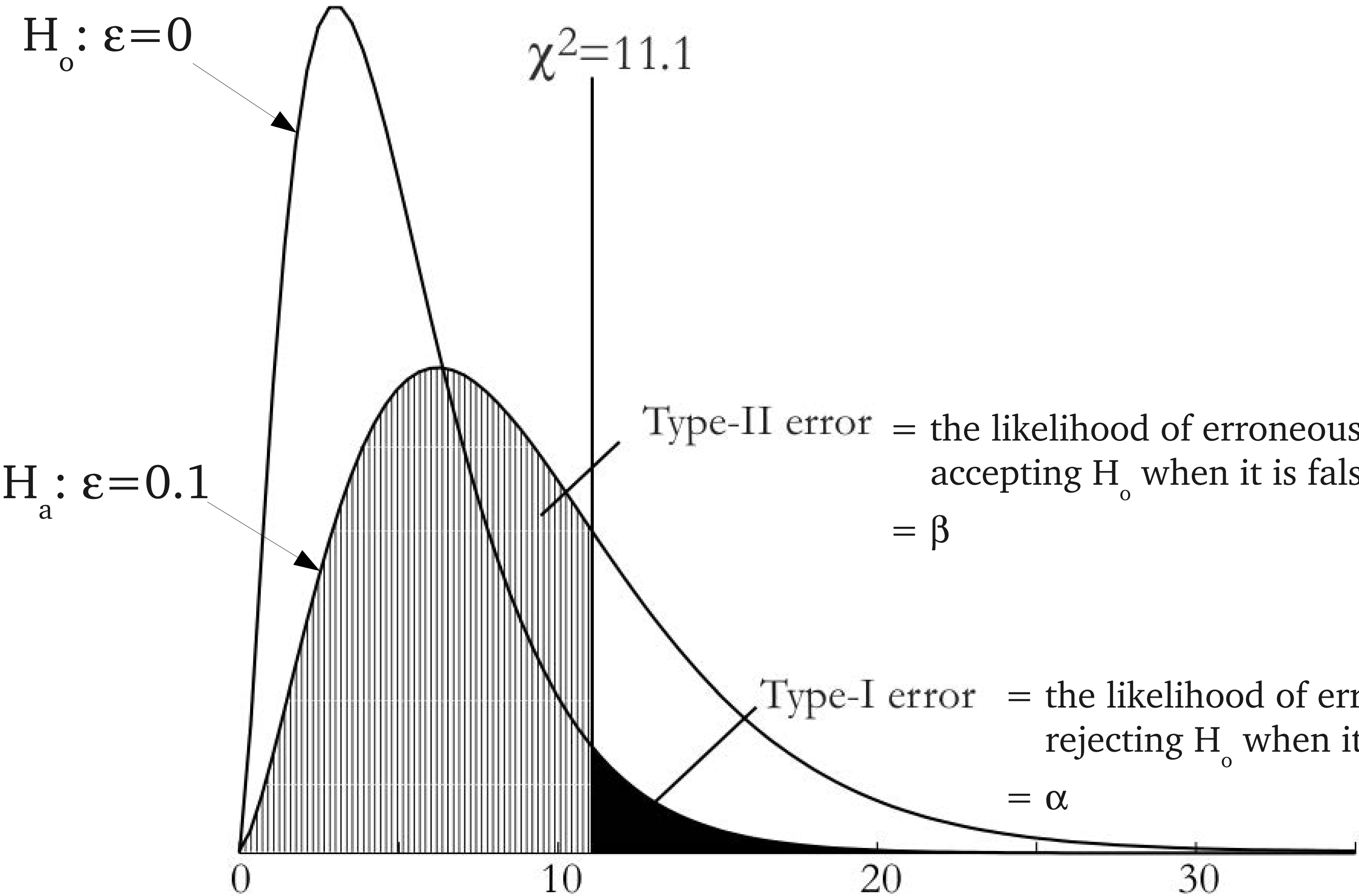
Both the QQ plots and the age spectra can become unwieldy if they contain more than a dozen or so samples. For example, Fig. 1 contains $n = 13$ kernel density estimates (KDEs, Vermeesch, 2012) showing the probability distributions of 2025 single grain age estimates, while the QQ-plots in Fig. 2 form an upper triangular matrix with $n(n-1)/2 = 78$ pairwise comparisons. This is simply too much information for the human eye to process. To solve this problem, we need a 'filter' removing the redundant features of the individual distributions while preserving and amplifying the significant differences between them. This paper makes the case that a standard statistical technique called multidimensional scaling (MDS) can be used effectively for this purpose (Sections 3 and 4).

In addition to the DZ ages, all but one (T) of the samples in the Chinese study were subjected to heavy mineral (HM) analysis. With the exception of samples 1 and 8, the HM analyses were performed on separate aliquots from the U–Pb measurements. For samples 1 and 8, the HM mounts were prepared by mixing leftover mineral separates from the DZ study. Between 201 and 419 grains were counted in the 63–250 μm size fraction of each sample, resulting in an additional 2901 datapoints. Part of the aim of this paper is to treat these categorical data on an equal footing with

* Tel.: +44 20 7679 2418.
E-mail address: p.vermeesch@ucl.ac.uk.

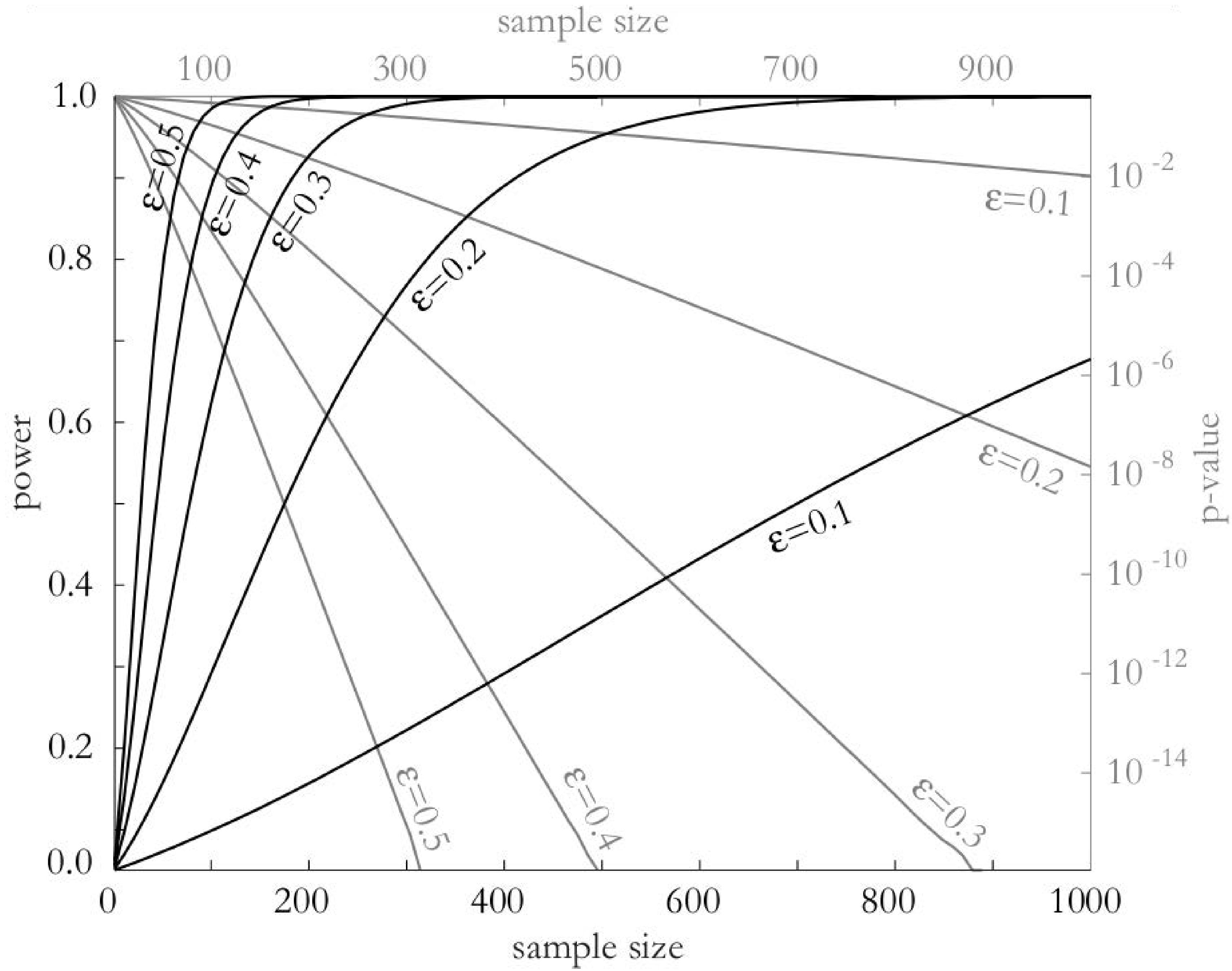
0009-2541/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemgeo.2013.01.010

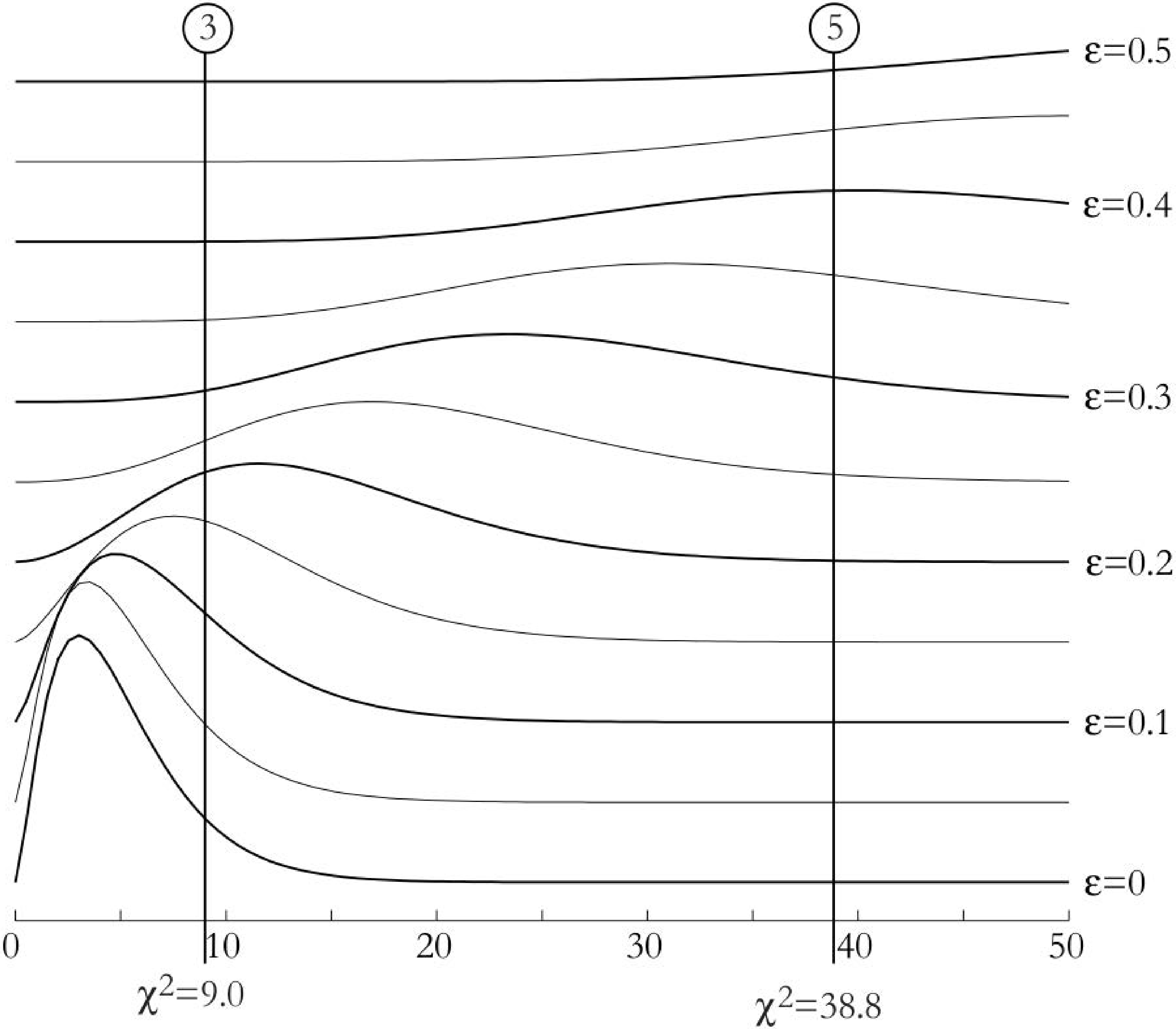
1. don't use p-values
2. do use effect sizes
3. MDS can make the interpretation of large datasets more objective
4. PCA/MDS treats petrographic data and geochronological data on an equal footing



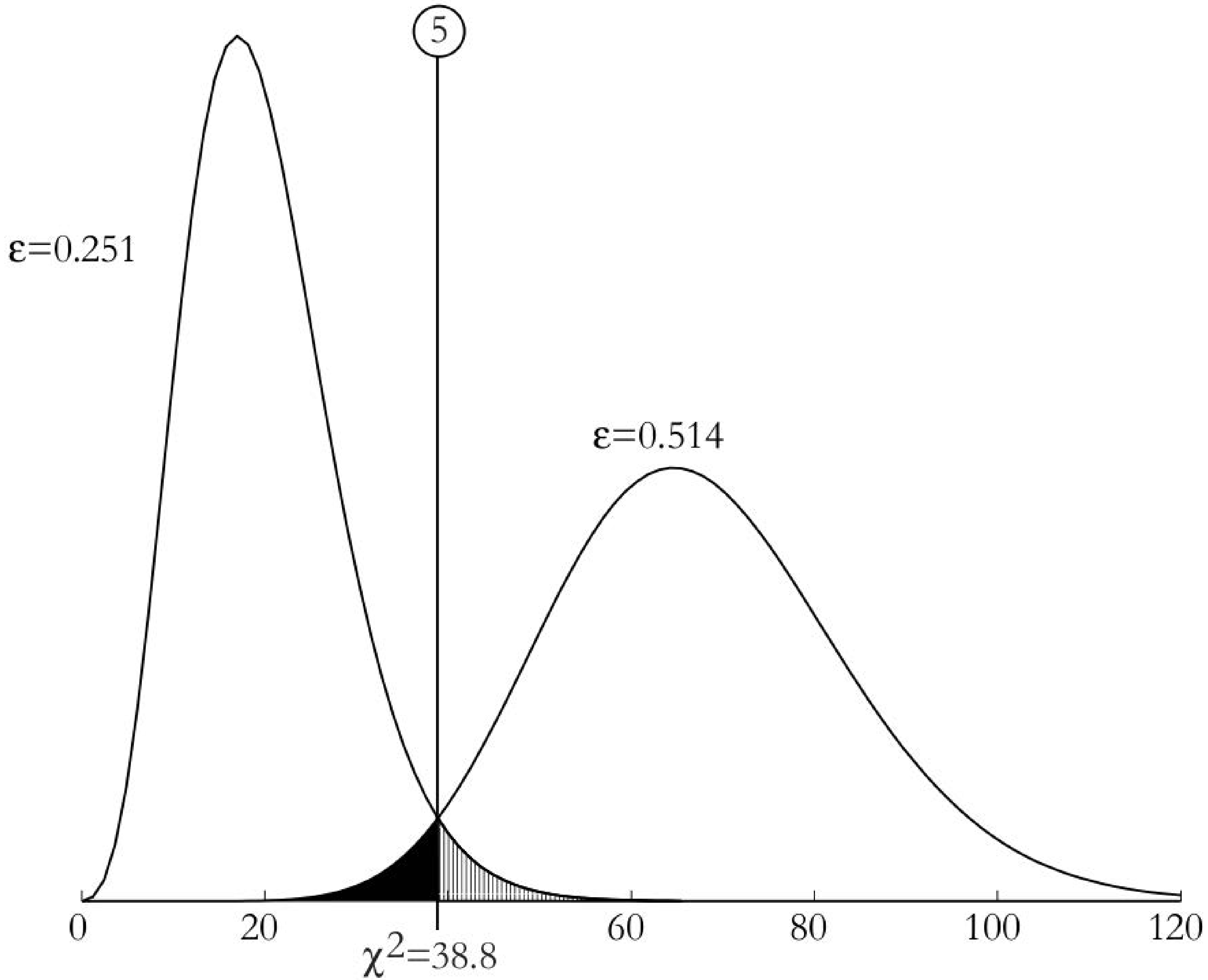
Type-II error = the likelihood of erroneously accepting H_o when it is false.
= β

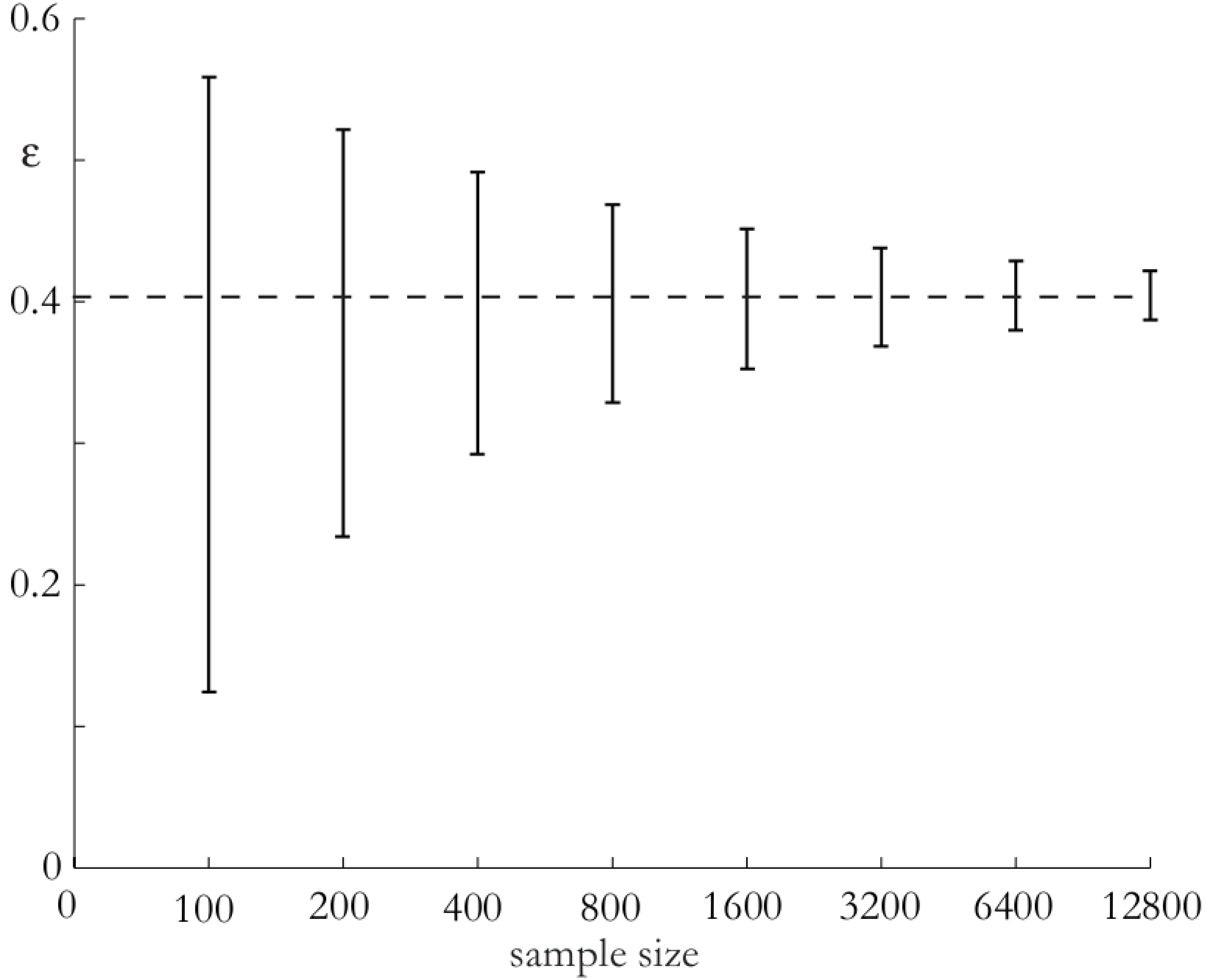
Type-I error = the likelihood of erroneously rejecting H_o when it is true.
= α


sample size
100
300
500
700
900
1.0
0.8
0.6
0.4
0.2
0.0
power
0
200
400
600
800
1000
sample size
10 -2
10 -4
10 -6
10 -8
10 -10
10 -12
10 -14
p-value
ε=0.5
ε=0.4
ε=0.3
ε=0.2
ε=0.1
ε=0.1
ε=0.2
ε=0.5
ε=0.4
ε=0.3


3
5
ε=0.5
ε=0.4
ε=0.3
ε=0.2
ε=0.1
ε=0
0
10
20
30
40
50
χ²=9.0
χ²=38.8

Confidence interval = all effect sizes >100α% likely to result in H_o being accepted




0.6
ε
0.4
0.2
0
0
100
200
400
800
1600
3200
6400
12800
sample size