#NewMR

A Presentation from
The Festival of NewMR – Training Day

3 December 2012

An Introduction to Latent Class Analysis for Marketing Segmentation

Tim Bock, Q

An Introduction to Latent Class Analysis for Marketing Segmentation

Tim Bock, Q

www.q-researchsoftware.com

tim.bock@q-researchsoftware.com

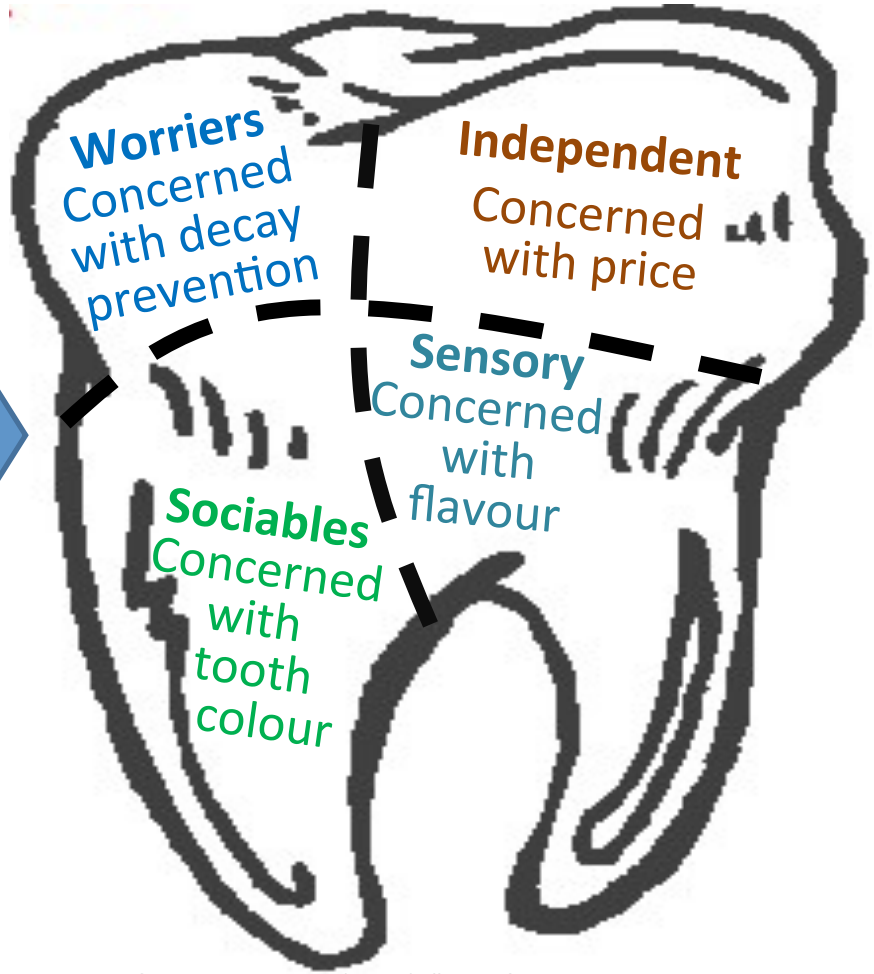
+61 425 241 989

Overview

- Latent class analysis versus cluster analysis
  - Theoretical difference: probabilities
  - Practical differences:
    - Non-numeric data (e.g., categorical data)
    - Missing values
- Application: what do research buyer's want?
  - Missing values
  - Response bias

Latent class analysis turns data into segments

StartDate_date	Q002	Q003	Q004	Q005	Q005_2	Q006	Q007__1
28/09/2012 09:47:34	1		1	2	2		1
28/09/2012 09:45:35	6		1	2	2		0
28/09/2012 09:56:17	1		2	2	2		1
28/09/2012 09:56:37	6		2	2	2		1
28/09/2012 09:55:19	11		3	2	2		1
28/09/2012 10:01:04	6		1	2	2		1
28/09/2012 10:16:46	3		1	1	1	5	1
28/09/2012 09:55:15	3		1	1	1	1	1
28/09/2012 10:59:20	1		1	2	2		1
28/09/2012 10:53:39	1		2	2	2		1
28/09/2012 11:01:35	3		1	2	2		0
28/09/2012 11:11:37	3		1	1	1	2	1
28/09/2012 11:00:27	6		1	2	2		0
28/09/2012 11:15:41	1		1	2	2		1
28/09/2012 11:35:34	1		2	2	2		0
28/09/2012 11:18:11	3		1	1	1	2	1
28/09/2012 11:28:24	6		1	2	2		1
28/09/2012 11:35:35	1		1	2	2		1
28/09/2012 11:40:34	1		2	2	2		1
28/09/2012 11:36:16	6		1	2	2		1
28/09/2012 11:53:12	15		2	2	2		1
28/09/2012 11:38:11	1		1	2	2		1
28/09/2012 11:38:10	1		1	2	2		1
28/09/2012 11:26:03	15		2	2	2		1



Adapted from: Haley, R. I. (1968). "Benefit Segmentation: A Decision Oriented Research Tool." Journal of Marketing **30**(July): 30-35.

Latent Class Analysis

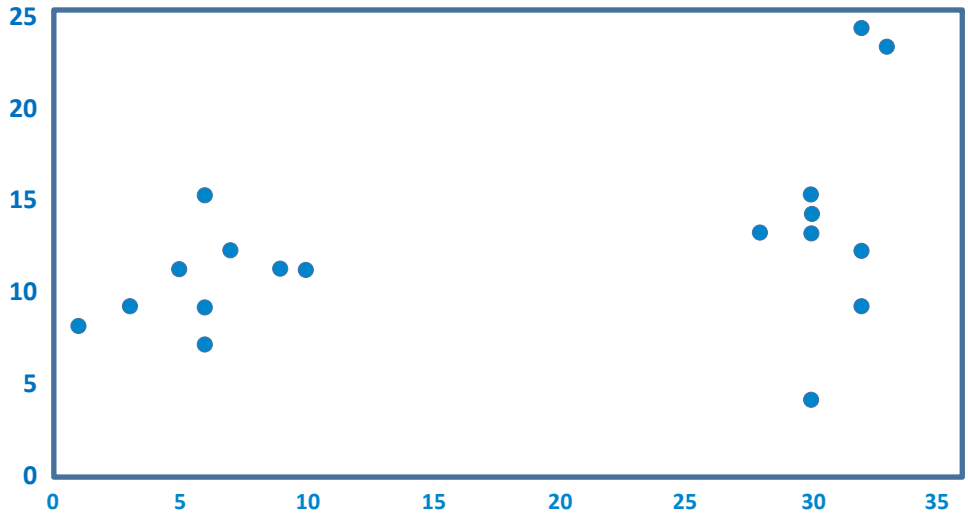
Cluster Analysis

Cluster Analysis versus Latent Class Analysis for segmentation

- Latent class analysis is theoretically superior
  - Clearly-stated assumptions
  - Cluster analysis is inconsistent with elementary laws of probability (in particular, *Bayes' Theorem)*
- Latent class analysis software is superior
  - Any type of data (via distributional assumptions): Categorical, Conjoint, Choice, MaxDiff, Rankings, etc.
  - "Mixed" data (e.g., categorical and numeric)
  - Missing values
  - Response biases

k-Means Cluster Analysis

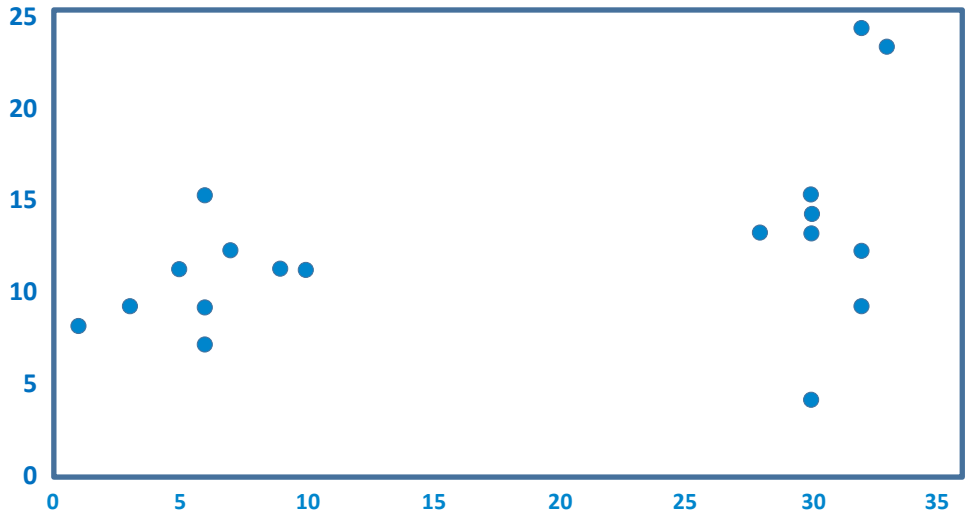
Specify number of clusters (*k*)



k-Means Cluster Analysis

Specify number of clusters (*k*)

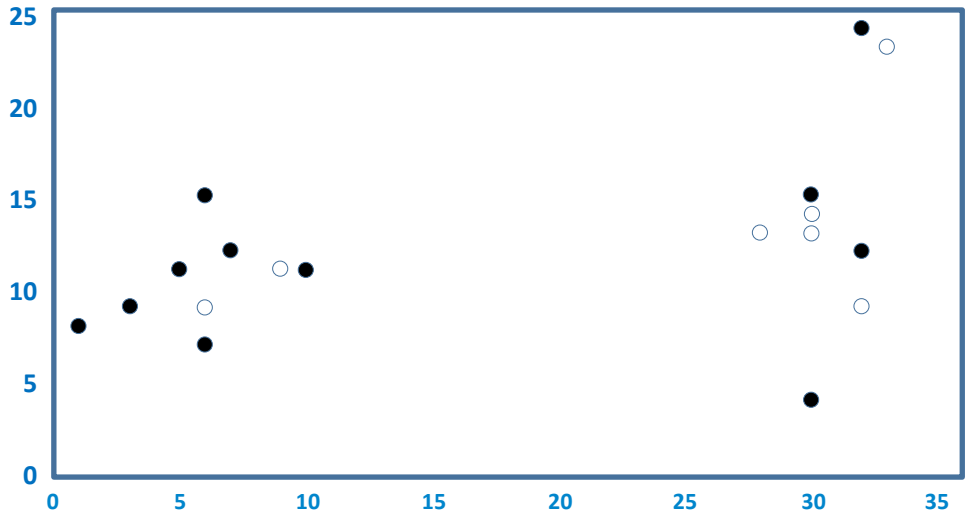
Randomly allocate respondents to clusters



k-Means Cluster Analysis

Specify number of clusters (*k*)

Randomly allocate respondents to clusters

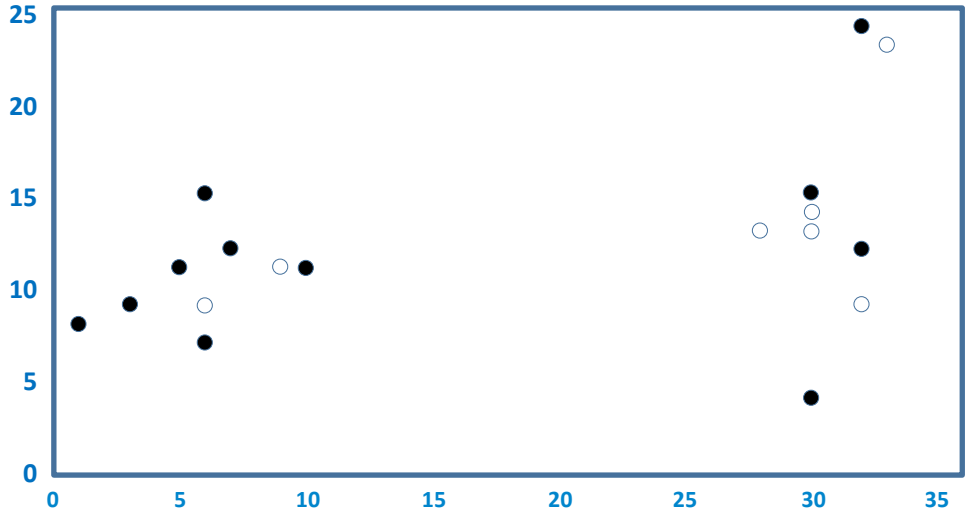


k-Means Cluster Analysis

Specify number of clusters (*k*)

Randomly allocate respondents to clusters

Compute cluster means

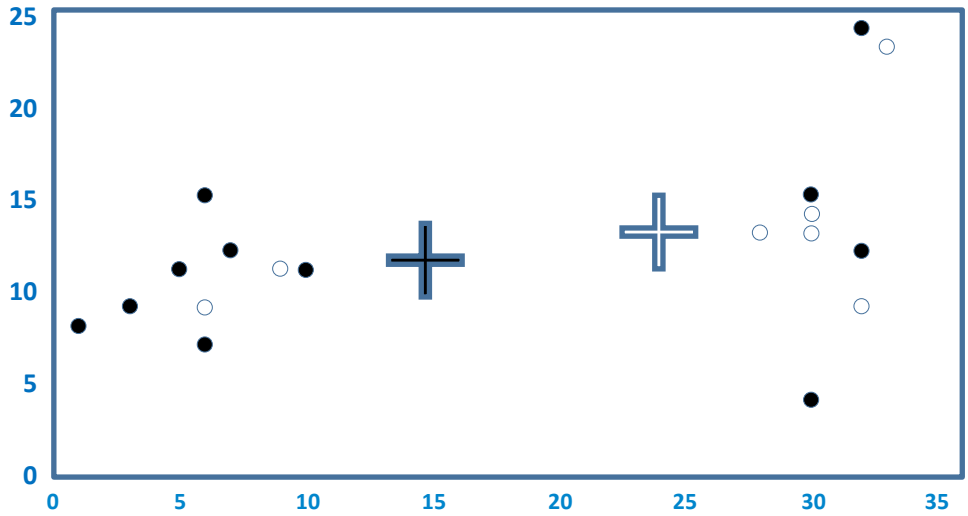


k-Means Cluster Analysis

Specify number of clusters (*k*)

Randomly allocate respondents to clusters

Compute cluster means



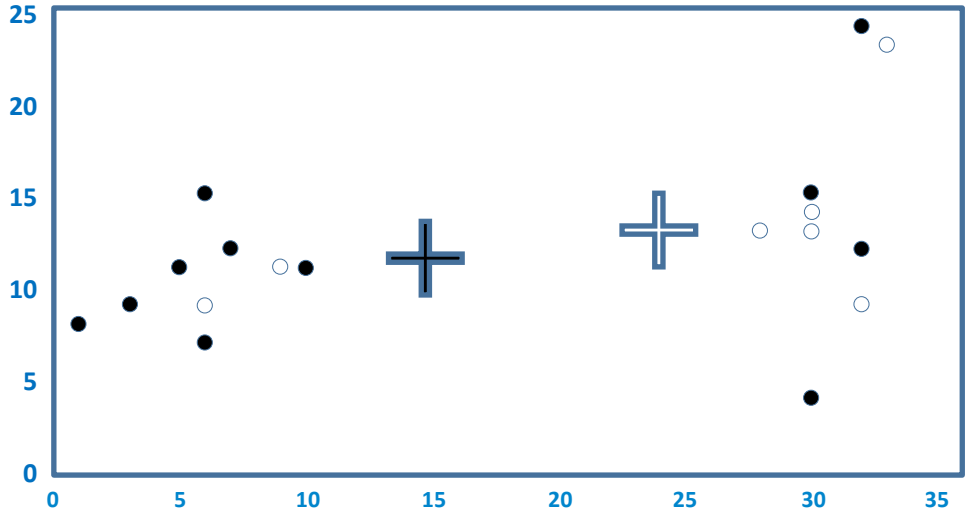
k-Means Cluster Analysis

Specify number of clusters (*k*)

Randomly allocate respondents to clusters

Compute cluster means

Allocate respondents to most similar clusters



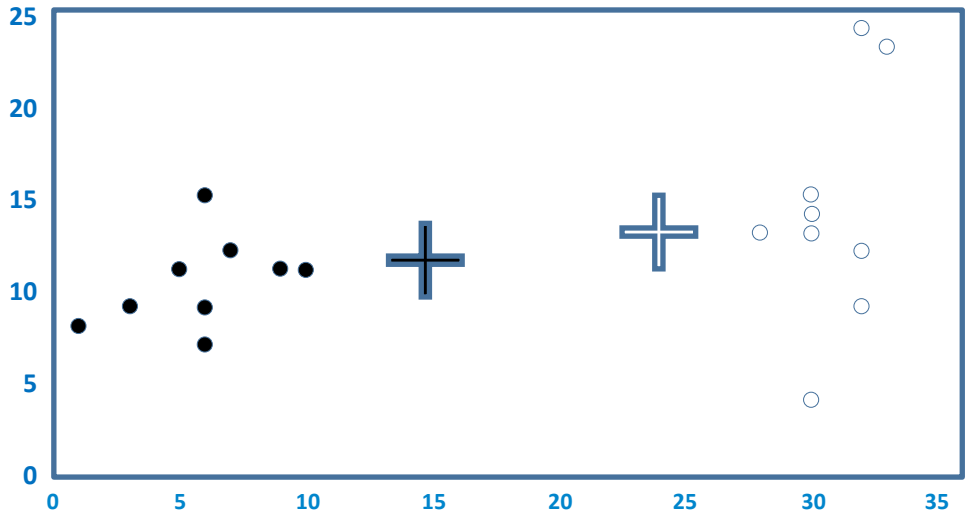
k-Means Cluster Analysis

Specify number of clusters (*k*)

Randomly allocate respondents to clusters

Compute cluster means

Allocate respondents to most similar clusters



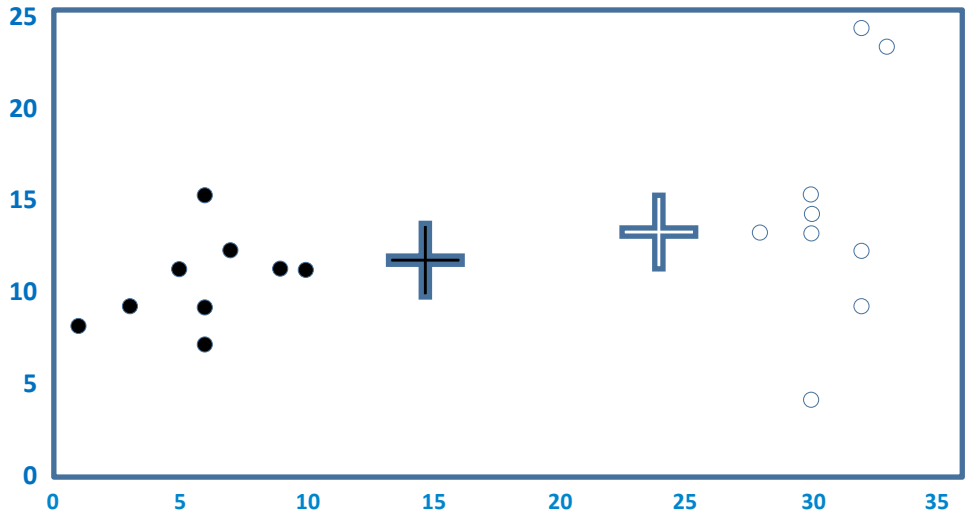
k-Means Cluster Analysis

Specify number of clusters (*k*)

Randomly allocate respondents to clusters

Compute cluster means

Allocate respondents to most similar clusters



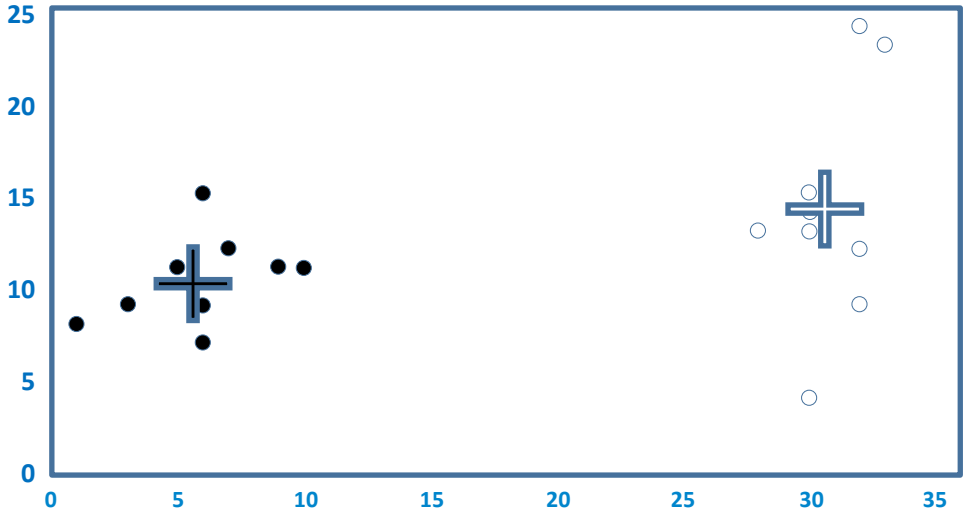
k-Means Cluster Analysis

Specify number of clusters (*k*)

Randomly allocate respondents to clusters

Compute cluster means

Allocate respondents to most similar clusters



k-Means Cluster Analysis

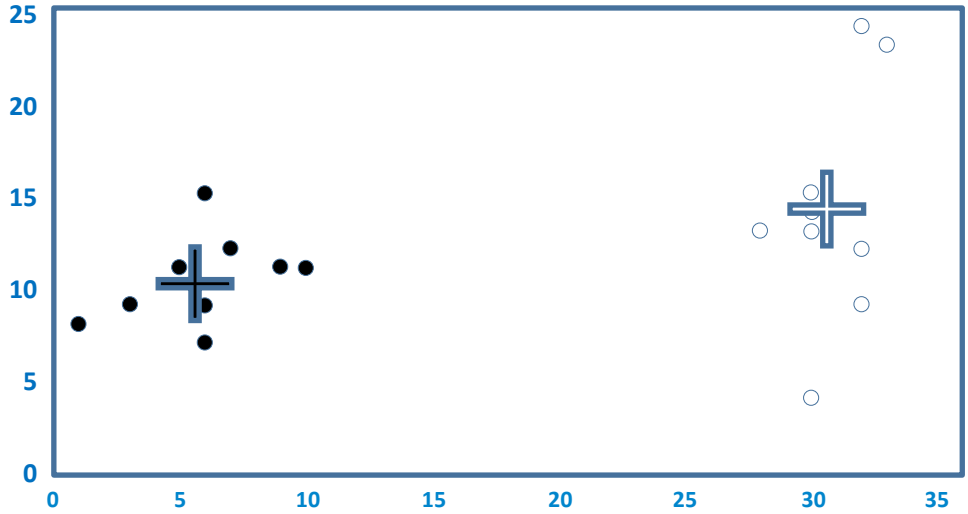
Specify number of clusters (*k*)

Randomly allocate respondents to clusters

Compute cluster means

Allocate respondents to most similar clusters

Repeat until changes in cluster means are small or non-existent



k-Means Cluster Analysis

Specify number of clusters (*k*)

Randomly allocate respondents to clusters

Compute cluster means

Allocate respondents to most similar clusters

Repeat until changes in cluster means are small or non-existent

Latent Class Analysis

Specify number of classes (*k*)

Randomly allocate respondents to classes

Compute class parameters*

Compute probability of each respondent being in each class

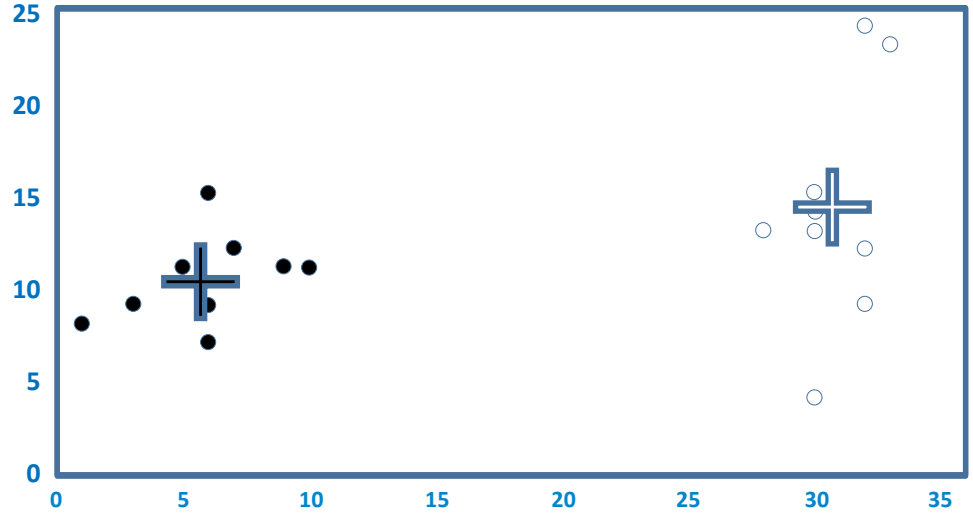
Repeat until changes in class parameters are small or non-existent

Allocate respondents classes with highest probabilities

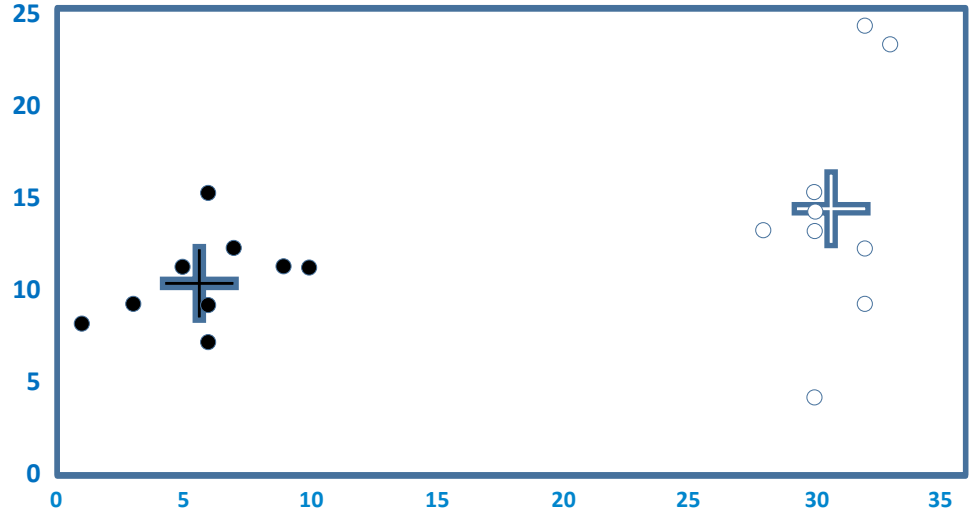
This is a comparison of batch k-means and Latent Class Analysis with an EM Algorithm. See Celeux and Govaert (1991), "Clustering criteria for discrete data and latent class models", Journal of Classification, 8(2) for a more mathematical comparison.

* The class parameters are computed as weighted averages of the segmentation variables, where the weights are the probabilities of each respondent being in each segment.

Cluster Analysis



Latent Class Analysis



Latent Class Analysis

Cluster Analysis

missing values

	A	B	C	D
1	1	2	3	4
2	1	2	3	4
3	4	3	2	1
4	4	3	2	1
5	1	2	.	1
6	.	2	2	1
7	1	2	2	.
8	1	.	2	1

How many clusters (or classes) can you see in this data?

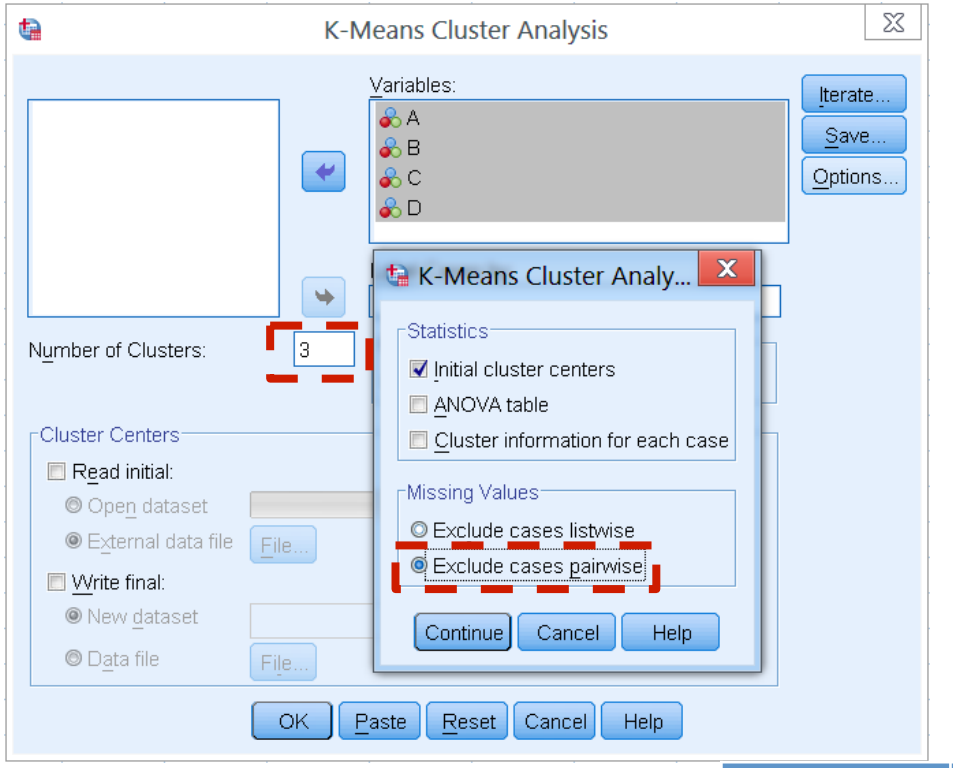
Missing values and latent class analysis

A	B	C	D	Classes
1	2	3	4	1
1	2	3	4	1
4	3	2	1	2
4	3	2	1	2
1	2		1	3
	2	2	1	3
1	2	2		3
1		2	1	3

Class means

	A	B	C	D
Cluster 1	1	2	3	4
Cluster 2	4	3	2	1
Cluster 3	1	2	2	1

Missing values and cluster analysis



	A	B	C	D	QCL_1
1	1	2	3	4	1
2	1	2	3	4	1
3	4	3	2	1	3
4	4	3	2	1	3
5	1	2	.	1	1
6	.	2	2	1	3
7	1	2	2	.	1
8	1	.	2	1	3

Cluster means

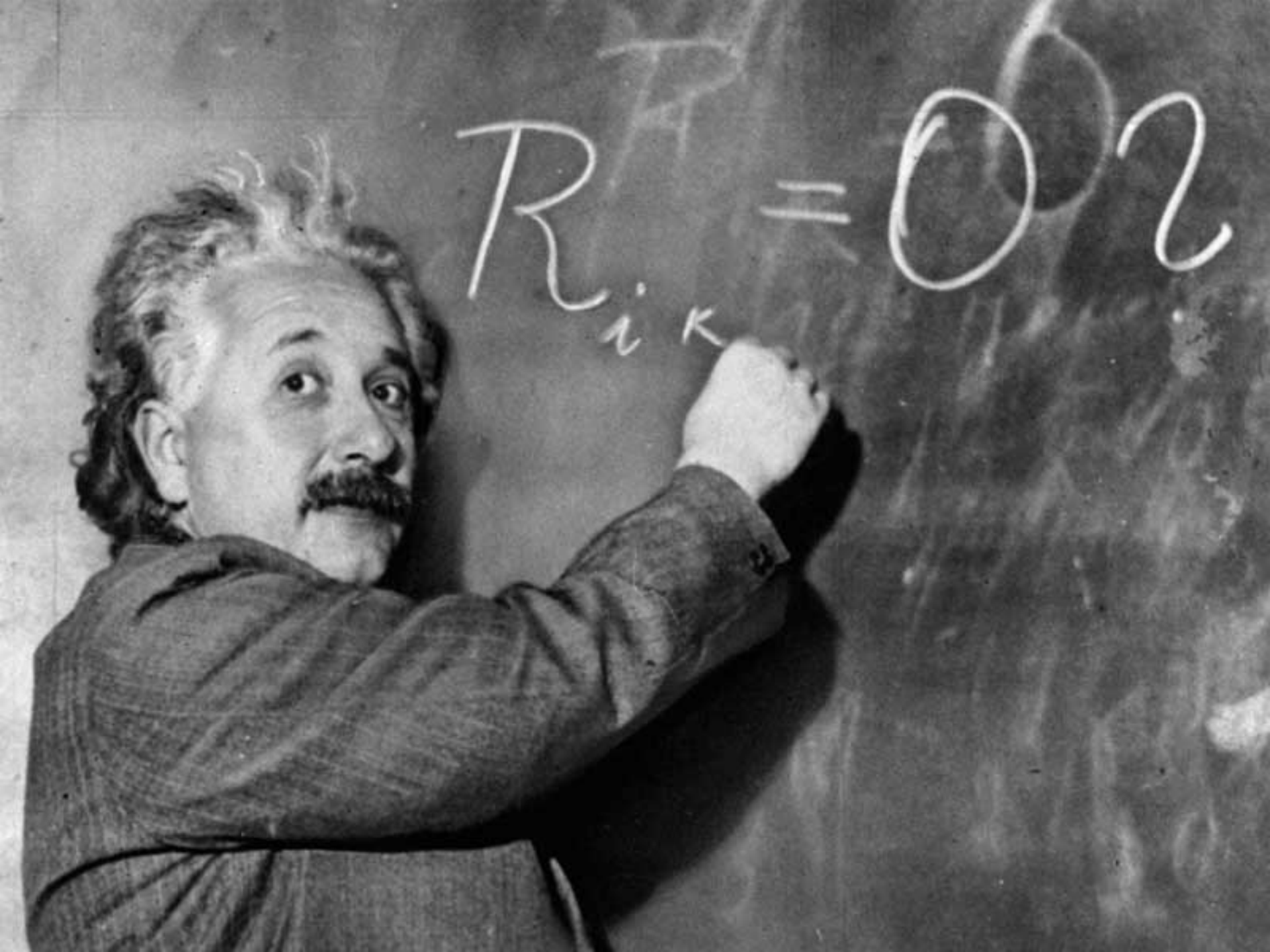
	A	B	C	D
Cluster 1	1	2	3	3
Cluster 2	MISSING	MISSING	MISSING	MISSING
Cluster 3	3	3	2	1

distributional assumptions

Distributional assumptions

	Variable		
ID	A	B	C
1	Agree	Agree	Neither
2	Agree	Disagree	Neither
3	Agree	Neither	Neither
4	Neither	Neither	Disagree

- Basic idea: instruct a latent class models about how to interpret the data
- Categorical assumption: look only at matches
  - Example: respondent *1* is most similar to *2* and *3* (i.e., they match on two variables)
- Numeric assumption: assign values and compute differences (e.g., Agree = 3, Neither = 2, Disagree = 1)
  - Example: respondent *1* is most similar to respondent *3*
- Ranking assumption: look at relative order
  - Respondent *1* is identical to respondent *4*


$R_{ik} = 0$

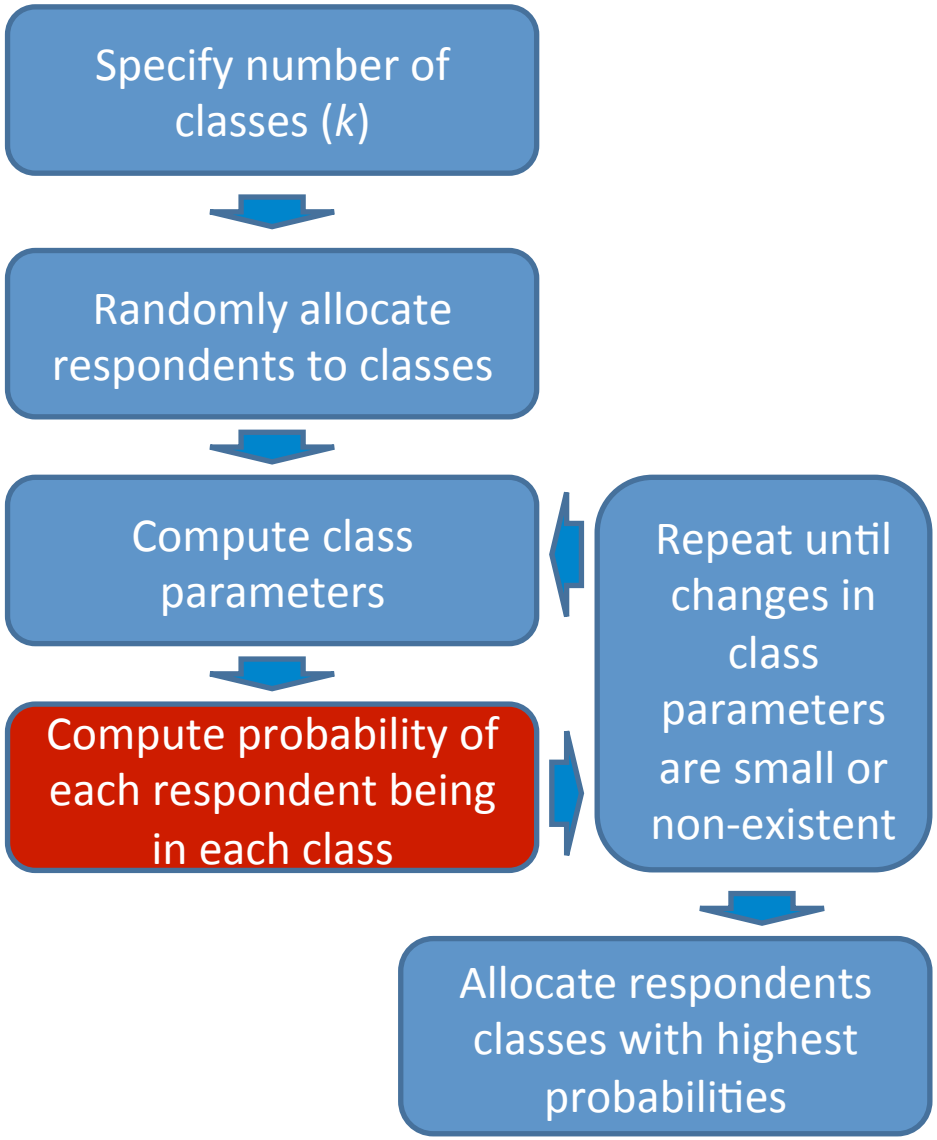
Example: Categorical data

Data

Shop	Agree (A) or disagree (D) that "It is important to shop around"
Understand	Agree (A) or disagree (D) that "I understand my company's communication needs"
Key	Agree (A) or disagree (D) that "Communications technology is key to our business"
Interested	Agree (A) or disagree (D) that "I am interested in communications technology"
Value	Agree (A) or disagree (D) that "Value for money is more important to us than getting the best technology"

ID	Shop	Understand	Key	Interest	Value
1	A	A	A	A	D
2	A	A	A	D	A
3	A	A	A	A	D
4	A	A	D	A	A
5	A	D	A	D	D
6	D	A	A	A	D
7	A	D	A	D	D
8	D	D	A	A	D
9	A	A	A	A	A
10	A	A	A	A	D
11	D	A	D	D	A
12	A	A	A	A	A
13	D	D	D	D	D
...	...	...	...	...	...

Latent Class Analysis



Data

ID	Shop	Understand	Key	Interest	Value
…	…	…	…	…	…
6	D	A	A	A	D
…	…	…	…	…	…

Parameters

	Size
Class 1	67%
Class 2	33%

	Shop	Under-stand	Key	Interest	Value
Agree	40%	40%	48%	16%	53%
Disagree	60%	60%	52%	84%	47%
Agree	65%	90%	88%	100%	26%
Disagree	35%	10%	12%	0%	73%

Looking at the **parameters**, which class do you think respondent 6 belongs to?

Computing the probability of each respondent being in each class

Data

ID	Shop	Understand	Key	Interest	Value
…	…	…	…	…	…
6	D	A	A	A	D
…	…	…	…	…	…

Parameters

	Size
Class 1	67%
Class 2	33%

	Shop	Under-stand	Key	Interest	Value
Agree	40%	40%	48%	16%	53%
Disagree	60%	60%	52%	84%	47%
Agree	65%	90%	88%	100%	26%
Disagree	35%	10%	12%	0%	73%

Initial best guess of probabilities is given by the class sizes:

Class 1: 67%
Class 2: 33%

Prior

Class conditional densities

Probability that somebody in each class would give answers:

Class 1: 60%×40%×48%×16%×47% = 1%

Class 2: 35%×90%×88%×100%×73% = 20%

$$\textit{Class 1:}\ \frac{67\%\times 1\%}{67\%\times 1\% + 3\%\times 20\%} = 9\%$$

$$\textit{Class 2:}\ \frac{33\%\times 20\%}{67\%\times 1\% + 33\%\times 20\%} = 91\%$$

Posterior probability
(Probability of being in a class)

Application

n = 1,145 market researchers (GRIT2012/2013)

"How important do you think each of the following attributes is to clients when they select a research provider?"

5 POINT SCALE

RANDOMLY SHOW 15 OF 25 ATTRIBUTES TO EACH RESPONDENT

Cluster Analysis

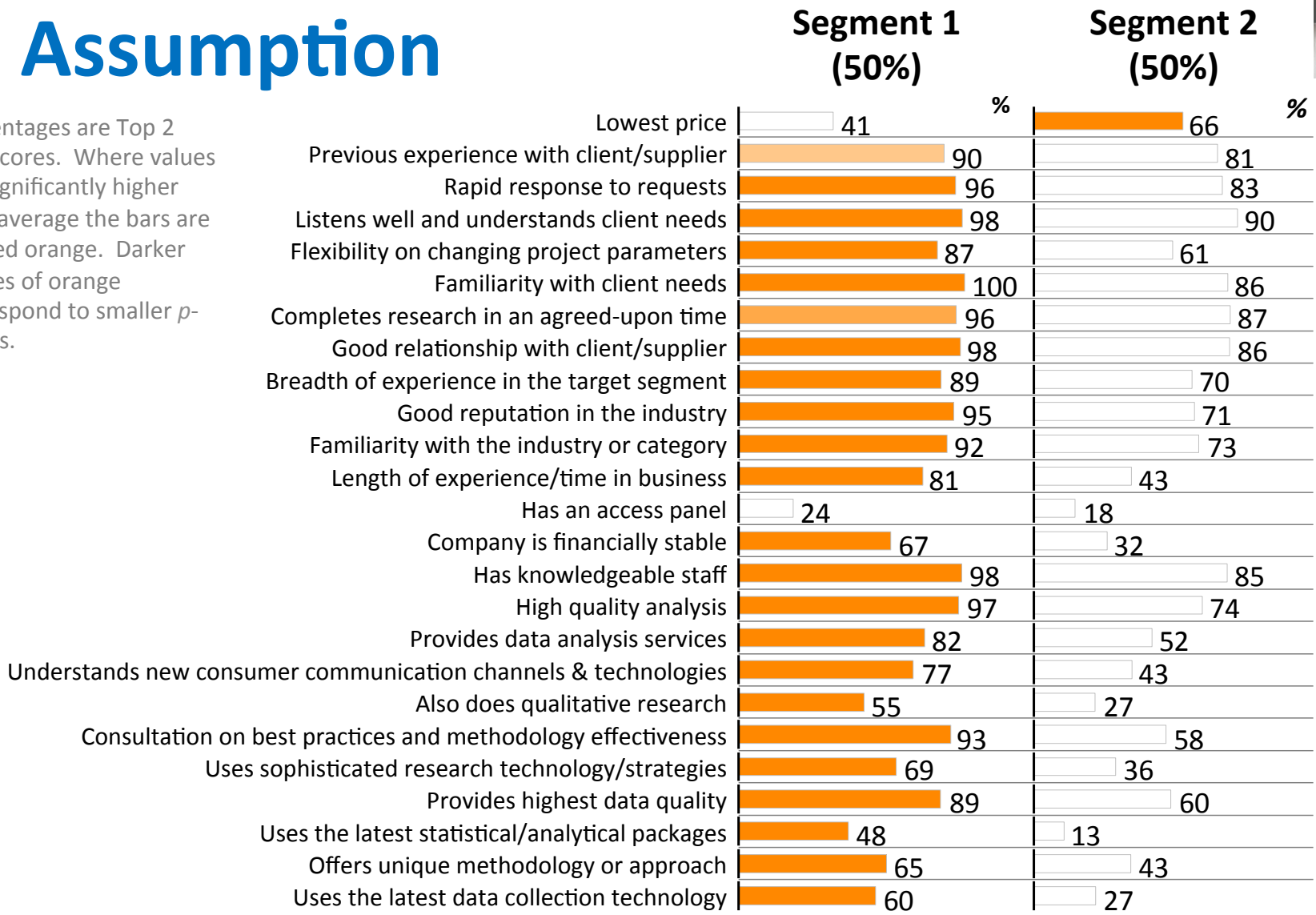
Numeric Assumption

Percentages are Top 2 Box Scores. Where values are significantly higher than average the bars are shaded orange. Darker shades of orange correspond to smaller *p*-values.

	Segment 1 (45%)	Segment 2 (11%)	Segment 3 (45%)
	%	%	%
Lowest price	50	68	55
Previous experience with client/supplier	88	73	87
Rapid response to requests	95	65	89
Listens well and understands client needs	98	67	97
Flexibility on changing project parameters	83	47	71
Familiarity with client needs	99	50	95
Completes research in an agreed-upon time	97	67	91
Good relationship with client/supplier	95	75	94
Breadth of experience in the target segment	90	31	81
Good reputation in the industry	93	40	82
Familiarity with the industry or category	92	33	85
Length of experience/time in business	86	16	51
Has an access panel	36	33	2
Company is financially stable	71	13	36
Has knowledgeable staff	97	58	96
High quality analysis	96	30	91
Provides data analysis services	86	28	59
Understands new consumer communication channels & technologies	84	27	45
Also does qualitative research	66	16	22
Consultation on best practices and methodology effectiveness	96	18	71
Uses sophisticated research technology/strategies	75	27	37
Provides highest data quality	96	15	69
Uses the latest statistical/analytical packages	57	10	7
Offers unique methodology or approach	75	31	39
Uses the latest data collection technology	79	17	14

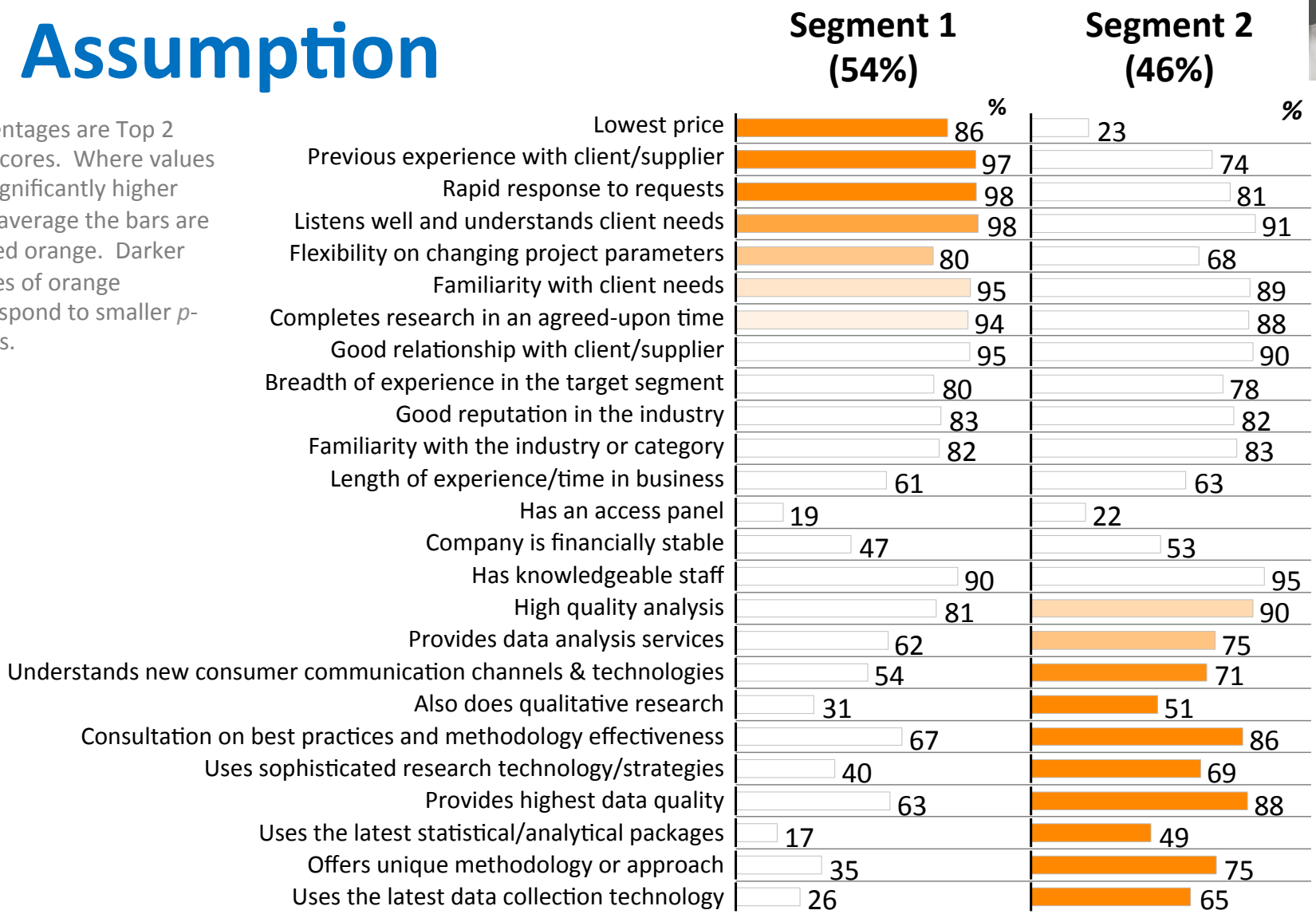
Categorical Assumption

Percentages are Top 2 Box Scores. Where values are significantly higher than average the bars are shaded orange. Darker shades of orange correspond to smaller *p*-values.



Ranking Assumption

Percentages are Top 2 Box Scores. Where values are significantly higher than average the bars are shaded orange. Darker shades of orange correspond to smaller *p*-values.

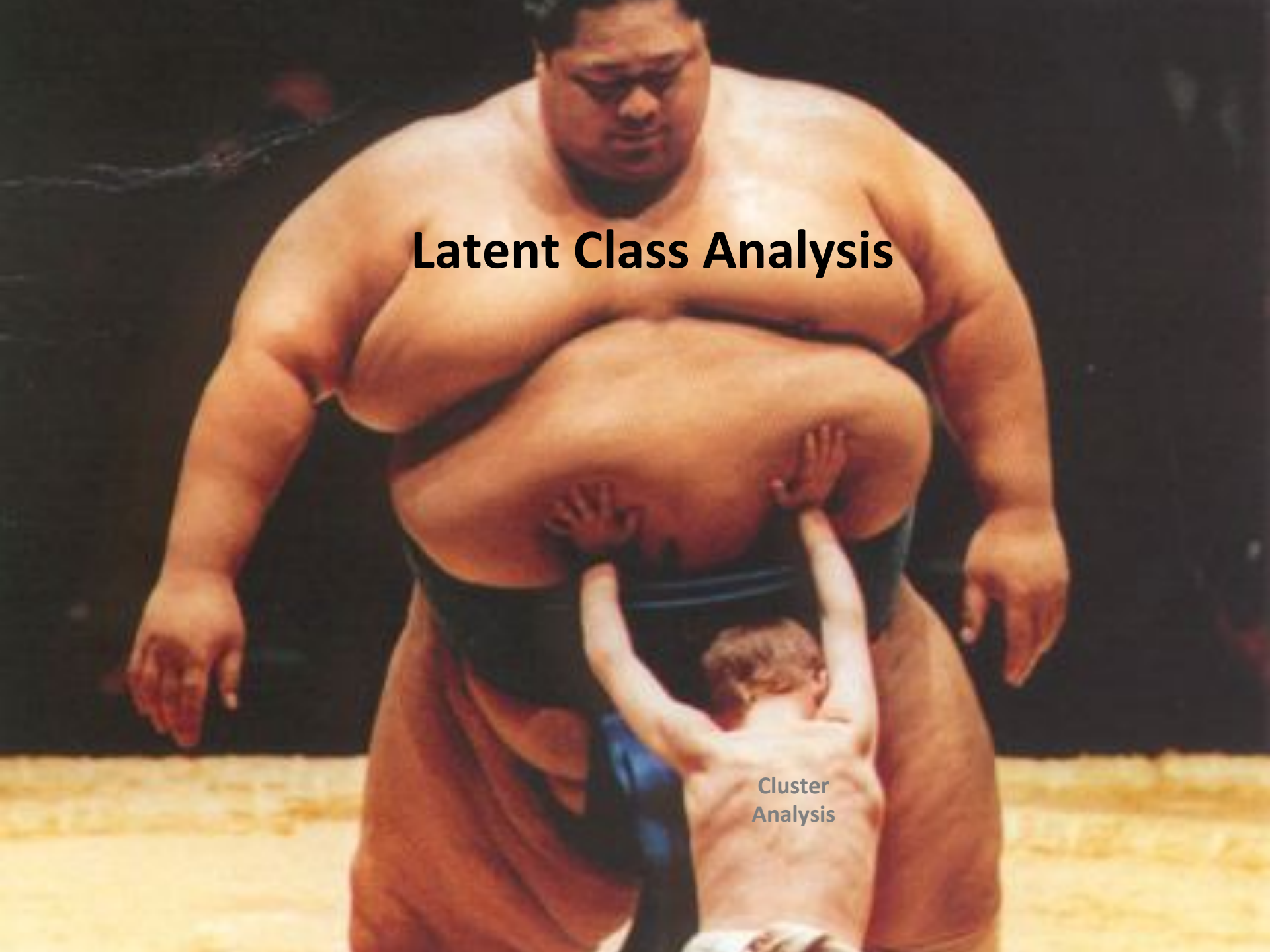


Latent class analysis software

Product	Data/distributional assumptions	Covariates*	Complex Sampling*
Sawtooth Software	Regression (discrete choice, ranks), Max-Diff	No	No
Q	Numeric, Binary, Categorical, Ranks, Partial Ranks, Ranks with Ties, Max-Diff, Regression (linear, discrete choice, ranks, partial ranks, ranks with ties, best-worst), Mixed data	No	No
Limdep	Regression (linear, discrete choice, censored, ranks, partial ranks, counts, survival, etc.)	Yes	No
SAS (PROC LCA/LTA/ Mixed)	Numeric, Binary, Categorical, Growth, Regression (discrete choice, ranks, partial ranks)	Yes	Yes
MPlus	Numeric, Binary, Categorical, Ordered, Categorical, Counts, Mixed data	Yes	Yes
Latent gold/Latent Gold Choice	Numeric, Binary, Categorical, Growth, Ranks, Partial Ranks, Counts, Regression (linear, discrete choice, censored, ranks, partial ranks)	Yes	Yes

* Covariates and the ability to handle complex sampling can be relevant when applying latent class analysis to non-segmentation problems (e.g., creating predictive models).

Latent Class Analysis



Thank you

Tim Bock

Q

Tim Bock, Q

www.q-researchsoftware.com
tim.bock@q-researchsoftware.com
+61 425 241 989