Exam 04/02/2020

Problem

Please open the file *Pisa_00_03_06_09 _12_15_18.sav.*

a) Please run a k-means cluster analysis with the three variables "'Lesekompetenz (Reading literacy)"', "'Mathematische Grundbildung (Mathematical literacy)"', "'Naturwissenschaftliche Grundbildung (Scientific literacy)"' for the competition in the year 2018. As the number of clusters please select the value four.

1. How many cases are in each cluster? Please complete the following table:

	Cluster			
	1	2	3	4
Number of cases				

2. Please complete the following table:

average value

	Cluster			
	1	2	3	4
Reading Literacy				
Mathematical Literacy				
Scientific Literacy				

3. Please comment the four clusters.
4. Which cluster does Germany belong to?
5. Which cluster does PSJZ China belong to?

b) Run a hierarchical cluster analysis with the three variables "'Lesekompetenz (Reading literacy)"', "'Mathematische Grundbildung (Mathematical literacy)"', "'Naturwissenschaftliche Grundbildung (Scientific literacy)"' for the competition in the year 2018.

1. How many clusters should be constructed? (Give reasons!)
2. Which cluster does Germany belong to?
3. Which cluster does PSJZ China belong to?

c) Please class the cases of the two variables "'Mathematische Grundbildung (Mathematical literacy)"' and "'Naturwissenschaftliche Grundbildung (Scientific literacy)"' for the competition in the year 2018 into three classes. The classes should have about the same number of cases. Please check with a level $\alpha = 0.05$ test, whether the two classed variables are stochastically independent.

1. What is the name of the test?

2. **Is the rule of thumb fulfilled? (Give reasons!)**
3. **How small is the p -value?**
4. **What is the test decision? (Comment!)**
5. **Compute a measure of association between the classed variables.Comment!**

d) **What is a measure to verify the loss of information if the data set is plotted in scatterplot, where x-axis and y-axis are the first and the second principal components of the three variables "'Lesekompetenz (Reading literacy)"', "'Mathematische Grundbildung (Mathematical literacy)"', "'Naturwissenschaftliche Grundbildung (Scientific literacy)"' for the competition in the year 2018? Compute and comment the value of this measure.**

Problem 31.01.2019

Please open the file *Credit_card.sav*.

a) Please run a hierarchical cluster analysis with the two variables ”items” and ”spent”.

1. Please complete the following table:

Stage	Coefficients
26 277	
26 278	
26 279	

2. How many clusters should be constructed? (Explain!)

b) Please run a K-Means cluster analysis with the two variables ”‘items”’ and ”‘spent”’. Three clusters should be constructed.

1. Please complete the following table:

average value

	Cluster		
	1	2	3
Number of items			
Amount spent			
Number of cases			

2. Please comment the three clusters.

c) Please run a level 0.05 test to check whether gender and clustermembership are dependent. Proceed as follows:

1. What is the name of the test?
2. Please check the rule of thumb of the test.
3. Please give the value of the p-value?
4. What is the test decision? (Comment!)
5. Please compute the measure of association Gamma between gender and clustermembership. What is the value of Gamma?

Technology Arts Sciences Cologne
Faculty of Economics, Business and Law
Prof. Dr. Arrenberg
Room 221, Tel. 39 14
jutta.arrenberg@th-koeln.de

Master: Quantitative Methods

Old Exams

Problem (25.01.2017)
Please open the file *telco_extra.sav*.

a) Please run a hierarchical cluster analysis with the five variables Month with service, Age in years, Years at current address, Household income in thousands, Number of people in household. How many clusters should be constructed? (Give reasons!)

b) Please run a *k*-means cluster analysis with the five variables Month with service, Age in years, Years at current address, Household income in thousands, Number of people in household. As the number of clusters please select the value three.

1. How many cases are in cluster 1 resp. 2 resp. 3?

	Cluster		
	1	2	3
Cases			

2. Please complete the following table:

Average Values

	Cluster		
	1	2	3
Month with service			
Age in years			
Years at current address			
Household income in thousands US-Dollar			
Number of people in household			

3. Please comment the three clusters.

4. Check with a statistical test whether the medians of the variable “Years with current employer” are the same across the three clusters. What is the name of the test? What are the assumptions of the test? What is the p-value of the test? What indicates this p-value?

5. What are the three values of the empirical medians of the variable "Years with current employer" in the clusters?

Empirical Medians

	Cluster		
	1	2	3
Years with current employer			

c) Suppose you will run an ordinal regression for the dependent variable Y="Calling card last month". What is the link function if the variable Y is transformed into an ordinal leveled variable with the three categories "0 up to 10", "more than 10 up to 20" and "more than 20"? (Give reasons!)

Exam QM 04/02/2020

Pisa Survey 2018

Number of Cases in each Cluster

Cluster	1	32,000
	2	9,000
	3	25,000
	4	11,000
Valid		77,000
Missing		7,000

Final Cluster Centers

	Cluster			
	1	2	3	4
Reading	485,06	363,78	414,40	522,64
Mathematical	488,72	367,56	417,36	537,36
Scientific	485,56	372,11	420,08	531,55

Cluster 1: second best performance
Cluster 2: poorest performance
Cluster 3: third best performance
Cluster 4: best performance

Germany belongs to cluster 1.

China belongs to cluster 4.

Stage	Coefficients
75	35.057
76	45.177

The greatest jump of the coefficients happens from stage 75 to stage 76.
Recommended number of clusters = N-75=77-75=2.

Germany belongs to cluster 1.

China belongs to cluster 2.

Percentile Group of Mathe_2018 * Percentile Group of Naturw_2018 Crosstabulation

Count

		Percentile Group of Naturw_2018			
		1	2	3	Total
Percentile Group of Mathe_2018	1	22	3	0	25
	2	4	19	4	27
	3	0	4	22	26
Total		26	26	26	78

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	82,519[a]	4	,000
Likelihood Ratio	86,807	4	,000
Linear-by-Linear Association	56,225	1	,000
N of Valid Cases	78		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 8,33.

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Gamma	,967	,017	18,509	,000
N of Valid Cases		78			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Classed Points in science and mathematics are dependent.

Gamma=0.967 positive strong relationship between mathematical literacy and scientific literacy. High points in mathematics are going along with high points in science.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,932	97,740	97,740	2,932	97,740	97,740	1,526	50,865	50,865
2	,052	1,745	99,485	,052	1,745	99,485	1,459	48,620	99,485
3	,015	,515	100,000						

Extraction Method: Principal Component Analysis.

99.485 % of the total variance is explained by the first two principal components.

Exam 31/01/2019

Credit_card.sav

Stage	Coefficient
26277	31.226
26278	75.160
26279	117.824

75.160 – 31.226 = 43.934

117.824 – 75.160 = 42.664

Number of Clusters = n – 26277=26280 – 26277 = 3 Cluster

Final Cluster Centers

	Cluster		
	1	2	3
Number of items	4	7	1
Amount spent	294,38	625,90	48,92

Cluster 1 = medium number of items, medium amount spent

Cluster 2 = highest number of items, highest amount spent

Cluster 3 = smallest number of items, smallest amount spent

Cluster Number of Case * Gender Crosstabulation

Count

		Gender		
		Male	Female	Total
Cluster Number of Case	1	4330	4089	8419
	2	1582	1547	3129
	3	7528	7204	14732
Total		13440	12840	26280

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	,718[a]	2	,698
Likelihood Ratio	,718	2	,698
Linear-by-Linear Association	,173	1	,678
N of Valid Cases	26280		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 1528,78.

p-value = 0.698 > 0.05
No rejection of the null hypothesis, i.e. Gender and Cluster Membership are stochastically independent.

QCL_ordered * Gender Crosstabulation

Count

		Gender		
		Male	Female	Total
QCL_ordered	low number items, low amount spent	7528	7204	14732
	medium no items, medium amount spent	4330	4089	8419
	high no items, high amount spent	1582	1547	3129
Total		13440	12840	26280

X= Gender (male, female) dichotomous variable

Y= Cluster membership (1=medium group, 2=top group, 3=poor group) nominal

Recoding Cluster membership into an ordinal leveled variable 1=poor group, 2=medium group, 3=top group) for to calculate gamma. Gamma = 0.001

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Gamma	,001	,011	,062	,951
N of Valid Cases		26280			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Telco_extra.sav **Exam QM 25/01/2017**

a) Hierarchical cluster analysis

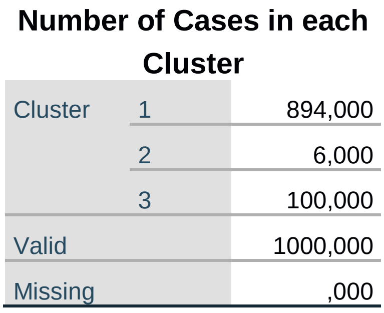
Greatest jump oft he coefficients

Stage	Coefficient
998	197.601
999	537.970

Number of cluster=n-998=1000-998=2 cluster

b) K-means cluster analysis

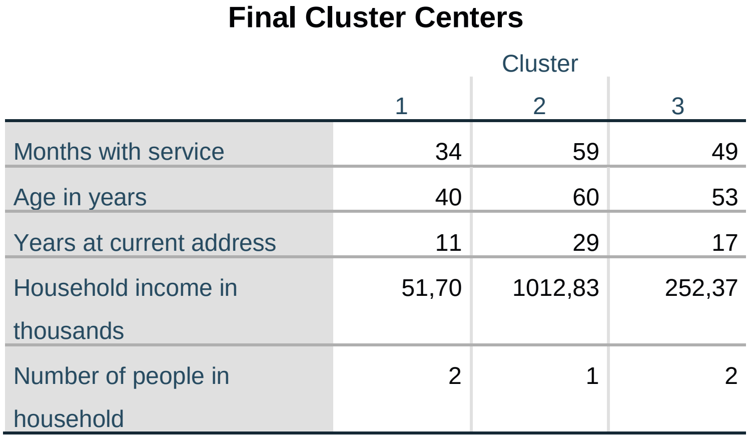
1.



Number of Cases in each Cluster

Cluster	1	894,000
	2	6,000
	3	100,000
Valid		1000,000
Missing		,000

2.



Final Cluster Centers

	Cluster		
	1	2	3
Months with service	34	59	49
Age in years	40	60	53
Years at current address	11	29	17
Household income in thousands	51,70	1012,83	252,37
Number of people in household	2	1	2

3. Cluster 1: youngest average age, shortest time of service and at current address, lowest income, average 2 people in household
   Cluster 2: oldest average age, longest time of service and at current address, highest income, single household
   Cluster 3: median average age, median time of service and at current address, median income, average 2 people in household

4. Kruskal-Wallis test
   Stochastic independence of „Years with current employer“ across the three cluster
   p-value =0.000
   Rejection of H0, at least two medians of „Years with current employer“ differ significantly across the three cluster

5.

Report

Median

Cluster Number of Case	Years with current employer
1	7,00
2	32,00
3	26,00
Total	8,00

c)

Class	Cases
≤10	434
10 - ≤20	308=742-434
>20	258

Link function: negative log log