Aufgabe (03.07.2018)

a) Öffnen Sie die Datei *Pisa_00_03_06_09 _12_15.sav*.

1. **Führen Sie eine hierarchische Clusteranalyse mit den drei Variablen „Lesekompetenz“, „Mathematische Grundbildung“, „Naturwissenschaftliche Grundbildung“ über alle sechs Austragungsjahre 2000, 2003, 2006, 2009, 2012, 2015 durch. Wie hoch ist die Anzahl der Cluster? (Begründung!)**
2. **Führen Sie eine Clusterzentrenanalyse (*k*-Means-Clusteranalyse) mit den drei Variablen „Lesekompetenz“, „Mathematische Grundbildung“, „Naturwissenschaftliche Grundbildung“ über alle sechs Austragungsjahre 2000, 2003, 2006, 2009, 2012, 2015 durch. Es sollen dabei genau drei Cluster gebildet werden.**
   - **Wie viele Fälle liegen in den einzelnen Clustern?**

	Cluster		
	1	2	3
Anzahl Fälle			

   - **Interpretieren Sie die drei Cluster.**
   - **Geben Sie nach Cluster getrennt die Länder an, die im Cluster mit den wenigsten Fällen und in dem Cluster mit den zweitwenigsten Fällen liegen.**
3. **Bilden Sie nur für das letzte Austragungsjahr 2015 sowohl für die Variable Mathematische Grundbildung als auch für die Variable Naturwissenschaftliche Grundbildung jeweils zwei etwa gleich stark besetzte Klassen. Prüfen Sie mit einem geeigneten Test zum Niveau $\alpha = 0{,}05$, ob die beiden klassierten Variablen stochastisch unabhängig sind.**
   - **Wie heißt der Test?**
   - **Ist die Faustregel erfüllt? (Begründung!)**
   - **Wie groß ist der *p*-Wert?**
   - **Wie lautet die Testentscheidung? (Interpretation!)**
   - **Liegt ggf. ein Zusammenhang zwischen den beiden klassierten Variablen vor? Falls ja, welcher?**

b) Mit welcher Maßzahl lässt sich der Informationsverlust bemessen, der bei einer Darstellung eines Datensatzes in einem Streudiagramm mit den beiden Achsen der ersten beiden Hauptkomponenten entsteht?

Lösung:

a) 1.

Schritt	Koeffizienten
21	77,531
22	229,865

höchster Sprung der Koeffizienten von Schritt 212 auf Schritt 22
♯ Cluster $= n - 21 = 23 - 21 = 2$ Cluster

2.

	Cluster		
	1	2	3
Anzahl Fälle	15	7	1

Cluster 1: Mittelfeld
Cluster 2: Spitzenreiter
Cluster 3: Schlusslicht
Cluster 1: Deutschland ...
Cluster 2: Japan, Finnland, Kanada, Südkorea, Neuseeland, Australien, Schweiz
Cluster 3: Mexiko

3.
- Chi-Quadrat-Unabhängigkeitstest
- minimale erwartete Häufigkeit = 17,5 $\geq$ 1 und keine Zelle hat eine erwartete Häufigkeit kleiner als fünf; d.h. die Faustregel ist erfüllt
- p-Wert ≈ 0;
- d.h. Ablehnung von H_0 ; d.h. Mathematische Grundbildung und Naturwissenschaftliche Grundbildung sind nicht stochastisch unabhängig
- $\gamma = 0{,}993$ d.h. geringe Punktzahl in mathematischer Grundbildung geht einher mit geringer Punktzahl in naturwissenschaftlicher Grundbildung. Und eine hohe Punktzahl in mathematischer Grundbildung geht einher mit einer hohen Punktzahl in naturwissenschaftlicher Grundbildung

b) Daran, wie viel Prozent der Gesamtvarianz durch die ersten beiden Hauptkomponenten erklärt wird.

Technische Hochschule Köln
Fakultät für Wirtschafts- und Rechtswissenschaften
Prof. Dr. Arrenberg
Raum 221, Tel. 39 14
jutta.arrenberg@th-koeln.de

Master: Quantitative Methoden

Alte Klausuren

Aufgabe (06.07.2016)
Öffnen Sie die Datei *employee_data.sav*.

a) Führen Sie ein hierarchische Clusteranalyse mit den drei Variablen Anfangsgehalt, Beschäftigungsdauer (in Monaten), Berufserfahrung (in Monaten) durch. Wie hoch ist die Anzahl der Cluster? (Begründung!)

b) Führen Sie eine Clusterzentrenanalyse (k-Means-Clusteranalyse) mit den drei Variablen Anfangsgehalt, Beschäftigungsdauer (in Monaten), Berufserfahrung (in Monaten) durch, wobei genau drei Cluster gebildet werden sollen.

1. Wie viele Fälle liegen in den einzelnen Clustern?

	Cluster		
	1	2	3
Anzahl Fälle			

2. Bitte füllen Sie die nachfolgende Tabelle aus:

Arithmetisches Mittel

	Cluster		
	1	2	3
Anfangsgehalt			
Beschäftigungsdauer			
Berufserfahrung			

3. Interpretieren Sie die drei Cluster.
4. Prüfen Sie mit einem Test, ob das mediane Gehalt in jedem der drei Cluster gleich hoch ist. Wie heißt der Test? Was sind die Voraussetzungen des Tests? Wie hoch ist der p-Wert des Tests? Welche Schlussfolgerng lässt sich aus dem p-Wert ziehen?
5. Wie hoch sind die empirischen Mediane der Variablen „Gehalt“ in den einzelnen Clustern?

Empirische Mediane

	Cluster		
	1	2	3
Gehalt			

c) Klassieren Sie die Variable „Gehalt“ in vier etwa gleich große Klassen. Berechnen Sie anschließend eine geeignete Maßzahl für den Zusammenhang zwischen dem klassierten Gehalt und der Variable „Art der Tätigkeit“. Wie heißt die Maßzahl für diesen Zusammenhang? Wie stark ist der Zusammenhang? (Begründung!)

Aufgabe (07.07.2015)
Öffnen Sie die Datei *dmdata.sav* aus dem Tutorial zu SPSS.

a) Bilden Sie drei Cluster, indem Sie mit den Variablen „age“, „years at current residence“ und „children“ eine Clusterzentrenanalyse durchführen.

1. Tragen Sie bitte die arithmetischen Mittel der Variablen in den jeweiligen Clustern sowie die Anzahl der Fälle in dem jeweiligen Cluster in die nachfolgende Tabelle ein:

Arithmetische Mittel

	Cluster		
	1	2	3
Age Years at current residence Children			
n			

2. Interpretieren Sie die drei Cluster.
3. Prüfen Sie mit einem statistischen Test, ob die theoretischen Mediane der Variablen „income category“ in den drei Clustern gleich sind. Wie heißt der Test? Was sind die Voraussetzungen des Tests? Wie hoch ist der p-Wert des Tests? Was lässt sich aus dem p-Wert folgern?
4. In welchem Cluster ist der höchste empirische Median und in welchem Cluster ist der kleinste empirische Median?

b) Betrachten Sie die beiden Variablen „children“ und „income category“.

1. Fassen Sie die Werte 4 und 5 Kinder der Variablen Children durch Umkodieren in einer gemeinsamen Klasse zusammen, während die übrigen Werte von „children“ unverändert beleiben. Welche Skalierung hat die umkodierte Variable?
2. Prüfen Sie mit einem Test, ob die umkodierte Variable und die Variable „income category“ stochastisch unabhängig sind. Sind die Test-Voraussetzungen erfüllt? (Begründung!) Und wie lautet der p-Wert? Was lässt sich aus dem p-Wert folgern?
3. Beantworten Sie anhand des Wertes von Gamma die Frage, ob eher Personen mit niedrigem Einkommen viele Kinder haben oder eher Personen mit hohem Einkommen.

Aufgabe (11.07.2014)
Öffnen Sie die Datei *grocery_coupons.sav* aus ILIAS.

I Welche Skalierungen haben die nachfolgenden Variablen:

1. hlthfood = Health food store = Reformhaus
2. size = Size of store = Geschäftsgröße
3. amtspent = Amount spent = Rechnungsbetrag (in US$)

II Wählen Sie von der Variablen „hlthfood“ (Reformhaus) nur die Nicht-Reformhäuser aus. Wie viele Nicht-Reformhäuser sind kleine bzw. mittelgroße bzw. große Geschäfte? Tragen Sie die Anzahlen in die nachfolgende Tabelle ein:

Größe	Anzahl
klein	
mittel	
groß	

a) Prüfen Sie anhand dieser Stichprobe mit einem Test, ob Rechnungsbeträge in allen drei Geschäftsgröße-Klassen in etwa gleich hoch sind.

1. Wie heißt der Test?
2. Überprüfen Sie die Voraussetzungen zum Testen!
3. Wie hoch ist der p-Wert?
4. Wie wird der p-Wert interpretiert?
5. In welcher Geschäftsgröße-Klasse sind die Rechnungsbeträge am größten? (Begründung!)

b) Führen Sie mit den Variablen

- size
- week
- amtspent

eine hierarchische Clusteranalyse durch. Wie viele Cluster sind zu bilden? (Begründung!)

Aufgabe (03.07.2013)

I Öffnen Sie die Datei *survey_sample.sav* aus ILIAS.

a) Wählen Sie von der Variablen „degree“ (Höchster Abschluss) nur die Fälle mit den Abschlüssen „Bachelor“ und „Universitätsabschluss“ aus. Wie viele Befragte haben einen Bachelor-Abschluss und wie viele Befragte haben einen Universitätsabschluss?

b) Klassieren Sie anschließend die Variable „rincome“ (Einkommen des Befragten) in die zwei Klassen „Einkommen bis 24999 $“ und „Einkommen 25000 $ oder mehr“. Wie viele Fälle liegen in der ersten Klasse und wie viele Fälle liegen in der zweiten Klasse?

c) Betrachten Sie die beiden Variablen aus a) und b). Überprüfen Sie mit einem Chi-Quadrat-Test, ob die beiden Variablen „Einkommen des Befragten (bis 24999 $ oder mindestens 25000 $)“ und „Höchster Abschluss (Bachelor oder Universitätsabschluss)“ stochastisch unabhängig voneinander sind.

1. Schreiben Sie auf, ob und wie die Faustregel erfüllt ist.
2. Geben Sie den p-Wert an.
3. Interpretieren Sie den p-Wert.

d) Wie stark ist der Zusammenhang zwischen den beiden Variablen „Einkommen des Befragten (bis 24999 $ oder mindestens 25000 $)“ und „Höchster Abschluss (Bachelor oder Universitätsabschluss)“ in der Stichprobe? Interpretieren Sie die Richtung und Stärke dieses Zusammenhangs.

II Wie verändert sich die Korrelation nach Bravais-Pearson für zwei metrische Variablen X=„Alter (in Jahren)“ und Y=„Höhe des Einkommens ins GE“, wenn alle Gehälter um 0,5 GE angehoben werden?

Aufgabe (11.07.2012)
Öffnen Sie die Datei *car_insurance_claims.sav* aus dem Tutorial von SPSS.

a) Welche Skalierungen haben die Variablen

   1. „Vehicle age“ (Alter des Fahrzeugs)?
   2. „Average cost of claims“ (durchschnittliche Kosten einer Schadensmeldung)?
   3. „Number of claims“ (Anzahl der Schadensmeldungen)?

b) Bilden Sie eine neue Variable „Kosten“, die sich als Produkt der beiden Variablen „Number of claims“ und „Average cost of claims“ ergibt. Klassieren Sie anschließend die Variable „Kosten“ in vier gleich stark besetzte Klassen. Wie viele Fälle liegen in jeder der vier Klassen?

c) Sind die beiden Variablen „klassierte Kosten“ und „Vehicle age“ stochastisch unabhängig? (Begründung!)

d) Wie stark ist der Zusammenhang zwischen den beiden Variablen „klassierte Kosten“ und „Vehicle age“? Interpretieren Sie die Richtung und Stärke dieses Zusammenhangs.

Aufgabe (08.07.2011)

Öffnen Sie die Datei *employee_data.sav* aus dem Tutorial von PASW.

a) Welche Skalierungen haben die Variablen

   1. „tätig“ (Art der Tätigkeit)?
   2. „mind“ (Minderheit)?
   3. „geschl“ (Geschlecht)?

b) Klassieren Sie die Variable „gehalt“ in vier etwa gleich stark besetzte Klassen. Wie viele Fälle liegen in jeder der

   1. ersten Klasse?
   2. zweiten Klasse?
   3. dritten Klasse?
   4. vierten Klasse?

c) Sind die beiden Variablen „klassiertes Gehalt“ und „Geschlecht“ stochastisch unabhängig? (Begründung!)

d) Wie stark ist der Zusammenhang zwischen den beiden Variablen „klassiertes Gehalt“ und „Geschlecht“? Interpretieren Sie die Richtung und Stärke dieses Zusammenhangs.

Klausur Master QM vom 06.07.2016

Employee_data.sav

Hierarchische Clusteranalyse

Schritt 472 Koeffizient 7.500,784

Schritt 473 Koeffizient 19.980,073

Anzahl Fälle minus größter Sprung der Koeffizienten=474-472=2 Cluster

Clusterzentrenanalyse mit drei Clustern

Clusterzentren der endgültigen Lösung

	Cluster		
	1	2	3
Anfangsgehalt	64.160	32.156	14.419
Beschäftigungsdauer	84	81	81
Berufserfahrung in Monaten	202	89	96

Bericht

Gehalt

Clusternummer des Falls	N	Median
1	3	103.500,00
2	61	65.000,00
3	410	27.450,00
Insgesamt	474	28.875,00

Percentile Group of gehalt * Art der Tätigkeit Kreuztabelle

Anzahl

		Art der Tätigkeit			
		Büro	Bewachung	Management	Gesamt
Percentile Group of gehalt	1	120	0	0	120
	2	115	2	0	117
	3	93	25	1	119
	4	35	0	83	118
Gesamt		363	27	84	474

Symmetrische Maße

		Wert	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Kontingenzkoeffizient	,658	,000
Anzahl der gültigen Fälle		474	

Hypothesentestübersicht

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Mediane von Gehalt sind über doe Kategorien von Clusternumme des Falls identisch.	Mediantest bei unabhängige n Stichprobe n	,000	Nullhypothese ablehnen

Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau i
05.

Lösungen der QM-Master-Klausuren

Lösungsvorschlag vom 07.07.2015

Clusterzentren der endgültigen Lösung

	Cluster		
	1	2	3
Age	45	59	30
Years at current residence	10	9	9
Children	1	2	0

Anzahl der Fälle in jedem Cluster

Cluster	1	4528,000
	2	2420,000
	3	3052,000
Gültig		10000,000
Fehlend		,000

Kruskal-Wallis-Test mit der Variablen income category und der

Gruppe=Clusterzugehörigkeit

Hypothesentestübersicht

	Nullhypothese	Test	Sig.	Entscheidung
1	Die Mediane von Income category (thousands) sind über doe Kategorien von Clusternummer des Falls identisch.	Mediantest bei unabhängigen Stichproben	,000	Nullhypothese ablehnen

Asymptotische Signifikanzen werden angezeigt. Das Signifikanzniveau i
05.

Empirische Mediane der Variablen income category in den drei Clustern:

Bericht

Income category (thousands)

Clusternummer des Falls	N	Median
1	4528	3,00
2	2420	4,00
3	3052	2,00
Insgesamt	10000	3,00

Income cat 1: <25

Income cat 2: 25 – 49

Income cat 3: 50 – 74

Income cat 4: 75 oder mehr

Income category (thousands) * child_neu Kreuztabelle

Anzahl

		child_neu					
		0	1	2	3	4 oder 5 Kinder	Gesamt
Income category (thousands)	<25	1491	148	28	0	0	1667
	25-49	1005	760	565	120	0	2450
	50-74	668	404	792	432	27	2323
	75+	1414	284	1346	499	17	3560
Gesamt		4578	1596	2731	1051	44	10000

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)
Chi-Quadrat nach Pearson	2508,202[a]	12	,000
Likelihood-Quotient	2797,976	12	,000
Zusammenhang linear-mit-linear	1213,176	1	,000
Anzahl der gültigen Fälle	10000		

a. 0 Zellen (0,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 7,33.

Symmetrische Maße

		Wert	Asymptotischer standardisierter Fehler[a]	Näherungsweises t[b]	Näherungsweise Signifikanz
Ordinal- bzgl. Ordinalmaß	Gamma	,398	,010	35,854	,000
Anzahl der gültigen Fälle		10000			

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

Lösungsvorschlag vom 11.07.2014

Health food store (0=no, 1=yes) binär

Size of store (small, medium, large) ordinal

Amount spent metrisch

n1=208, n2=520, n3=576

Kruskal-Wallis-Test

Stochastische Unabhängigkeit lässt sich mit SPSS nicht überprüfen

Ordinal oder metrische skalierte Variable

p-Wert=0,002 d.h. mindestens in zwei der drei Geschäftsgröße-Klassen sind die medianen Rechnungsbeträge signifikant unterschiedlich

Bericht

Amount spent

Size of store	Mittelwert	N	Median
Small	102,6879	208	97,8700
Medium	101,7483	520	101,4950
Large	95,7415	576	94,8950
Insgesamt	99,2449	1304	97,5900

d.h. in kleineren Geschäften wird mehr gekauft als in großen Geschäften

Schritt 1299 Koeffizient 9,9 und Schritt 1300 Koeffizient 16,919 ist der größte Sprung der Koeffizienten. Anzahl Fälle = 1304

1304-1299=fünf Cluster

Lösungsvorschlag vom 03.07.2013

Höchster Abschluss

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Niedriger als High School	430	15,2	15,2	15,2
	High School	1500	53,0	53,2	68,4
	Junior College	209	7,4	7,4	75,8
	Bachelor	478	16,9	16,9	92,7
	Universitätsabschluss	205	7,2	7,3	100,0
	Gesamt	2822	99,6	100,0	
Fehlend	KA	10	,4		
Gesamt		2832	100,0		

rincome_class

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	bis 24999	148	21,7	28,5	28,5
	25000 oder mehr	371	54,3	71,5	100,0
	Gesamt	519	76,0	100,0	
Fehlend	System	164	24,0		
Gesamt		683	100,0		

Höchster Abschluss * rincome_class Kreuztabelle

Anzahl

		rincome_class		
		bis 24999	25000 oder mehr	Gesamt
Höchster Abschluss	Bachelor	117	246	363
	Universitätsabschluss	31	125	156
Gesamt		148	371	519

Symmetrische Maße

		Wert	Asymptotischer standardisierter Fehler[a]	Näherungsweises t[b]	Näherungsweise Signifikanz
Ordinal- bzgl. Ordinalmaß	Gamma	,315	,104	3,048	,002
	Korrelation nach Spearman	,126	,041	2,877	,004[c]
Intervall- bzgl. Intervallmaß	Pearson-R	,126	,041	2,877	,004[c]
Anzahl der gültigen Fälle		519			

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

c. Basierend auf normaler Näherung

Symmetrische Maße

		Wert	Asymptotischer standardisierter Fehler[a]	Näherungsweises t[b]	Näherungsweise Signifikanz
Ordinal- bzgl. Ordinalmaß	Gamma	,315	,104	3,048	,002
Anzahl der gültigen Fälle		519			

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

$$r(X,Y+0,5) = r(X,Y)$$

Lösungsvorschlag vom 11.07.2012

Vehicle age (klassiert, also ordinal)

Average cost of claims (metrisch)

Number of claims (metrisch)

Percentile Group of Kosten

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	32	25,0	25,0	25,0
	2	32	25,0	25,0	50,0
	3	32	25,0	25,0	75,0
	4	32	25,0	25,0	100,0
	Gesamt	128	100,0	100,0	

Percentile Group of Kosten * Vehicle age Kreuztabelle

Anzahl

		Vehicle age				
		0-3	4-7	8-9	10+	Gesamt
Percentile Group of Kosten	1	0	1	10	21	32
	2	5	4	13	10	32
	3	9	13	9	1	32
	4	18	14	0	0	32
Gesamt		32	32	32	32	128

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)

Chi-Quadrat nach Pearson	85,000[a]	9	,000
Likelihood-Quotient	105,245	9	,000
Zusammenhang linear-mit-linear	69,076	1	,000
Anzahl der gültigen Fälle	128		

a. 0 Zellen (0,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 8,00.

Symmetrische Maße

		Wert	Asymptotischer standardisierter Fehler[a]	Näherungsweises t[b]	Näherungsweise Signifikanz
Ordinal- bzgl. Ordinalmaß	Gamma	-,795	,041	-16,240	,000
Anzahl der gültigen Fälle		128			

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

Lösungsvorschlag vom 08.07.2011

Art der Tätigkeit (nominal)

Minderheit (binär)

Geschlecht (dichotom)

Gehaltsklasse

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	bis 24000	120	25,3	25,3	25,3
	24001 bis 28800	117	24,7	24,7	50,0
	28801 bis 36600	118	24,9	24,9	74,9
	36601 oder mehr	119	25,1	25,1	100,0
	Gesamt	474	100,0	100,0	

Gehaltsklasse * Geschlecht Kreuztabelle

Anzahl

		Geschlecht		
		Männlich	Weiblich	Gesamt
Gehaltsklasse	bis 24000	16	104	120
	24001 bis 28800	57	60	117
	28801 bis 36600	82	36	118
	36601 oder mehr	103	16	119
Gesamt		258	216	474

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)
Chi-Quadrat nach Pearson	143,553[a]	3	,000
Likelihood-Quotient	157,895	3	,000
Anzahl der gültigen Fälle	474		

a. 0 Zellen (0,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 53,32.

Symmetrische Maße

		Wert	Asymptotischer standardisierter Fehler[a]	Näherungsweises t[b]	Näherungsweise Signifikanz
Ordinal- bzgl. Ordinalmaß	Gamma	-,736	,038	-15,494	,000
Anzahl der gültigen Fälle		474			

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.