Aufgabe (QM-Klausur vom 19.07.2021)
Bitte öffnen Sie die Datei *worldsales.sav*. In dieser Datei befinden sich folgende Variablen:

- Continent mit den sechs Kategorien Africa, Asia, Australia, Europe, North America, South America
- Product mit den drei Kategorien Product A, Product B, Product C
- Revenue (Umsatz) mit Werten im Intervall [153,62 ; 1 489,23]

Geben Sie bitte bei allen statistischen Tests, die Sie nutzen, die Voraussetzungen sowie die beiden Hypothesen an.

a)
1. Ist der Umsatz in den sechs Kontinent-Kategorien normalverteilt? Überprüfen Sie dies mit einem Test zum Niveau 0,05.
2. Sind die theoretischen Varianzen des Umsatzes in den sechs Kontinent-Kategorien homo- oder heterogen? Überprüfen Sie dies mit einem Test zum Niveau 0,05.
3. Ist der mediane Umsatz in den sechs Kontinent-Kategorien in etwa gleich hoch? Überprüfen Sie dies mit einem Test zum Niveau 0,05.

b) Betrachten Sie als Fälle lediglich den Kontinent Europa.

1. Ist der Umsatz von Product A in Europa normalverteilt? Überprüfen Sie dies mit einem Test zum Niveau 0,05.
2. Ist der Umsatz von Product B in Europa normalverteilt? Überprüfen Sie dies mit einem Test zum Niveau 0,05.
3. Ist der Umsatz von Product C in Europa normalverteilt? Überprüfen Sie dies mit einem Test zum Niveau 0,05.
4. Sind in Europa die theoretischen Varianzen des Umsatzes in den drei Produkt-Kategorien homo- oder heterogen? Überprüfen Sie dies mit einem Test zum Niveau 0,05.
5. Sind in Europa die mittleren Umsätze der drei Produkte in etwa gleich? Überprüfen Sie dies mit einem Test zum Niveau 0,05.

c) Betrachten Sie wieder alle Fälle.

1. Bitte führen Sie nur für die Variable „Revenue“ eine hierarchische Clusteranalyse durch. Wie viele Cluster sollten gebildet werden? (Begründung!)
2. Bitte führen Sie nur für die Variable „Revenue“ mit der empfohlenen Clusteranzahl der hierarchischen Clusteranalyse jetzt eine K-Means-Clusteranalyse durch und geben Sie die arithmetischen Mittel vom Umsatz in den jeweiligen Clustern an. Interpretieren Sie die erhaltenen Cluster.

Klausur QM am 19.07.2021

Datei worldsales.sav

a.1) Voraussetzung für den NV-Test ist die metrische Skalierung der zu testenden Variablen.
H0: Der Umsatz in einem Kontinent ist normalverteilt
H1: Der Umsatz in einem Kontinent ist nicht normalverteilt

Tests auf Normalverteilung

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	C_num	Statistik	df	Signifikanz	Statistik	df	Signifikanz
Revenue	Africa	,043	167	,200*	,990	167	,291
	Asia	,060	186	,200*	,990	186	,219
	Australia	,052	148	,200*	,992	148	,615
	Europe	,036	173	,200*	,997	173	,963
	North America	,039	167	,200*	,992	167	,535
	South America	,053	159	,200*	,991	159	,397

*. Dies ist eine untere Grenze der echten Signifikanz.

a. Signifikanzkorrektur nach Lilliefors

Da alle zwölf p-Werte größer als 0,05 sind, kann in allen sechs Kontinenten angenommen werden, dass der Umsatz normalverteilt ist.

a.2) Voraussetzung für den Levene-Test sind metrische Skalierung und Normalverteilung der Variablen Umsatz sowie stochastische Unabhängigkeit der Umsätze in den verschiedenen Kontinenten.

H0: Alle sechs theoretischen Varianzen sind gleich groß.

H1: Nicht alle sechs theoretischen Varianzen sind gleich groß.

Tests der Varianzhomogenität

		Levene-Statistik	df1	df2	Sig.
Revenue	Basiert auf dem Mittelwert	24,109	5	994	,000
	Basiert auf dem Median	23,789	5	994	,000
	Basierend auf dem Median und mit angepaßten df	23,789	5	809,139	,000
	Basiert auf dem getrimmten Mittel	24,112	5	994	,000

Da der p-Wert vom Levene-Test ca. null beträgt, liegt Heterogenität der Varianzen vor; d.h. mindestens zwei der sechs Varianzen sind signifikant unterschiedlich.

a.3) Voraussetzung für den Kruskal-Wallis-Test ist, dass die zu testende Variable ordinal oder metrisch skaliert ist und dass die Variable „Umsatz“ über die verschiedenen Gruppen (hier: Kontinente) stochastisch unabhängig ist.

H0: Die medianen Umsätze in den sechs Kontinenten sind gleich groß.

H1: Die medianen Umsätze in den sechs Kontinenten sind nicht gleich groß.

Teststatistiken[a,b]

	Revenue
Kruskal-Wallis-H	646,497
Df	5
Asymp. Sig.	,000

a. Kruskal-Wallis-Test

b. Gruppenvariable: C_num

Der p-Wert des Kruskal-Wallis-Test beträgt in etwa null, d.h. mindestens zwei der sechs medianen Umsätze sind signifikant unterschiedlich.

b.1) bis b.3)

Voraussetzung: metrische Skalierung

H0: Der Umsatz eines bestimmten Produkts ist normalverteilt.

H1: Der Umsatz des Produkts ist nicht normalverteilt.

Tests auf Normalverteilung

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	P_num	Statistik	df	Signifikanz	Statistik	df	Signifikanz
Revenue	Product A	,063	63	,200*	,987	63	,724
	Product B	,056	71	,200*	,988	71	,738
	Product C	,096	39	,200*	,960	39	,179

*. Dies ist eine untere Grenze der echten Signifikanz.

a. Signifikanzkorrektur nach Lilliefors

Da alle sechs p-Werte größer als 0,05 sind, ist der Umsatz für jedes der drei Produkte normalverteilt.

b.4) Voraussetzung für den Levene-Test: metrische Skalierung, NV, stochastische Unabhängigkeit.
H0: Alle drei Varianzen sind gleich groß.
H1: Die drei Varianzen sind nicht gleich groß.

Tests der Varianzhomogenität

		Levene-Statistik	df1	df2	Sig.
Revenue	Basiert auf dem Mittelwert	,980	2	170	,378
	Basiert auf dem Median	,952	2	170	,388
	Basierend auf dem Median und mit angepaßten df	,952	2	169,320	,388
	Basiert auf dem getrimmten Mittel	,964	2	170	,383

p-Wert Levene-Test = 0,378; d.h. Homogeniät der drei Varianzen.

b.5) Voraussetzungen der ANOVA: metrische Skalierung, NV, Homogenität der Varianzen, stochastische Unabhängigkeit.

H0: Die drei mittleren Umsätze von Produkt A,B,C sind gleich groß.

H1: Die drei mittleren Umsätze von Produkt A,B,C sind nicht gleich groß.

ANOVA

Revenue

	Quadratsumme	df	Mittel der Quadrate	F	Sig.
Zwischen den Gruppen	3661,303	2	1830,651	,061	,941
Innerhalb der Gruppen	5128739,995	170	30169,059		
Gesamt	5132401,298	172			

p-Wert ANOVA = 0,941; d.h. die mittleren Umsätze der drei Produkte A,B,C sind gleich groß.

c.1) größter Sprung der Koeffizienten von 25,737 in Schritt 997 auf 60,460 in Schritt 998.

Schritt	Koeffizient
997	25,737
998	60,460

Anzahl Cluster = n – 997 = 1000 – 997 = 3

c.2)

Clusterzentren der endgültigen Lösung

	Cluster		
	1	2	3
Revenue	810,71	499,35	1127,02

Cluster 1: mittelhoher Umsatz

Cluster 2: niedrigster Umsatz

Cluster 3: höchster Umsatz

Aufgabe (30 Punkte)
Bitte öffnen Sie die Datei *bankloan_binning.sav*.

a) **Berechnen Sie alle Hebelwerte des linearen Regressionsmodells:**

$$\text{Other debt} \approx b_0 + b_1 \cdot \text{Household income}$$

1. **Wie hoch ist das beobachtete Haushaltseinkommen des Datenpunkts mit dem größten Hebelwert?**
2. **Löschen Sie für die Berechnungen in den Teilaufgaben b) und c) den Datenpunkt mit dem größten Hebelwert aus dem Datensatz.**

b) **Führen Sie mit drei Variablen Ihrer Wahl sowohl eine hierarchische Clusteranalyse als auch eine *k*-Means Clusteranalyse durch.**

1. **Wie viele Cluster sind zu bilden? (Begründung!)**
2. **Interpretieren Sie die erhaltenen Cluster Ihrer *k*-Means Clusteranalyse.**

c) **Klassieren Sie die Variable age = „Alter“ in drei Altersklassen.**

1. **Wie viele Fälle liegen jeweils in einer Altersklasse?**
2. **Prüfen Sie mit einem Test zum Niveau $\alpha = 0{,}05$, ob die Variable mit der *k*-Means-Clusterzugehörigkeit aus Teilaufgabe b) und die Variable Altersklasse stochastisch unabhängig sind.**
3. **Berechnen und interpretieren Sie ein geeignetes Assoziationsmaß für die beiden Variablen *k*-Means-Clusterzugehörigkeit aus Teilaufgabe b) und Altersklasse.**

Aufgabe (30 Punkte)
Bitte öffnen Sie die Datei *bankloan_binning.sav.*

a) Berechnen Sie alle Hebelwerte des linearen Regressionsmodells:

Other debt $\approx b_0 + b_1 \cdot$ Household income

1. Wie hoch ist das beobachtete Haushaltseinkommen des Datenpunkts mit dem größten Hebelwert?

income	debtinc	creddebt	othdebt	default	PRE_1	RES_1	COO_1	LEV_1
2461,70	22,59	139,58	416,52	1	267,63983	148,87759	939,93646	,43456
433,00	4,25	4,49	13,91	0	45,42514	-31,51285	,35683	,01107
414,20	5,90	10,63	13,81	0	43,36587	-29,55851	,28398	,01002
393,00	7,11	15,03	12,91	0	41,04372	-28,13437	,22842	,00889
391,30	7,58	5,31	24,35	0	40,85751	-16,50620	,07785	,00881

Das beobachtete Haushaltseinkommen beträgt beim größten Hebelwert gleich 2461,70.

2. Löschen Sie für die Berechnungen in den Teilaufgaben b) und c) den Datenpunkt mit dem größten Hebelwert aus dem Datensatz.

b) Führen Sie mit drei Variablen Ihrer Wahl sowohl eine hierarchische Clusteranalyse als auch eine *k*-Means Clusteranalyse durch.

Gewählte Variablen:
- Household income in thousands
- Debt to income ratio (x100)
- Credit card debt in thousands

1. Wie viele Cluster sind zu bilden? (Begründung!)

Hierarchische Clusteranalyse:

Zusammenfassung der Fallverarbeitung[a,b]

Fälle					
Gültig		Fehlend		Gesamt	
N	Prozent	N	Prozent	N	Prozent
4999	100,0	0	,0	4999	100,0

a. Euklidisches Distanzmaß verwendet

b. Single Linkage

Fallzahl N = 4999

Schritt	Zusammengeführte Cluster		Koeffizienten	Erstes Vorkommen des Clusters		Nächster Schritt
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
4995	6	16	21,638	4994	4991	4997
4996	1	3	21,686	4992	4979	4998
4997	6	15	29,631	4995	0	4998
4998	1	6	48,495	4996	4997	0

Schritt 4997: Koeffizient = 29,631
Schritt 4998: Koeffizient = 48,495
Der größte Sprung der Koeffizienten liegt zwischen Schritt 4997 und 4998.

Empfohlene Anzahl an Clustern =N – 4997 = 4999 – 4997 = 2.
Es sind zwei Cluster zu bilden.

2. Interpretieren Sie die erhaltenen Cluster Ihrer *k*-Means Clusteranalyse.

Clusterzentren der endgültigen Lösung

	Cluster 1	Cluster 2
Household income in thousands	141,04	37,02
Debt to income ratio (x100)	10,22	10,06
Credit card debt in thousands	4,83	1,25

Cluster 1: Höchstes Haushaltseinkommen (in Tausend), höchstes Verhältnis Schulden/Einkommen (x100), höchste Kreditkartenschulden (in Tausend)

Cluster 2: Niedrigstes Haushaltseinkommen (in Tausend), niedrigstes Verhältnis Schulden/Einkommen (x100), niedrigste Kreditkartenschulden (in Tausend)

c) Klassieren Sie die Variable age = "Alter" in drei Altersklassen.

1. Wie viele Fälle liegen jeweils in einer Altersklasse?

Percentile Group of age

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	1768	35,4	35,4	35,4
	2	1498	30,0	30,0	65,3
	3	1733	34,7	34,7	100,0
	Gesamt	4999	100,0	100,0	

Klasse 1 (20 – 31): 1768 Fälle
Klasse 2 (32 – 38): 1498 Fälle
Klasse 3 (39 – 58): 1733 Fälle
Gesamt: 4999 Fälle

2. Prüfen Sie mit einem Test zum Niveau α= 0,05, ob die Variable mit der *k*-Means-Clusterzugehörigkeit aus Teilaufgabe b) und die Variable Altersklasse stochastisch unabhängig sind.

Pearson Chi-Quadrat-Test
Faustregel erfüllt?

Clusternummer des Falls * Percentile Group of age Kreuztabelle

Anzahl

		Percentile Group of age 1	2	3	Gesamt
Clusternummer des Falls	1	10	73	406	489
	2	1758	1425	1327	4510
Gesamt		1768	1498	1733	4999

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)
Pearson-Chi-Quadrat	576,721[a]	2	<,001
Likelihood-Quotient	608,198	2	<,001
Zusammenhang linear-mit-linear	516,297	1	<,001
Anzahl der gültigen Fälle	4999		

a. 0 Zellen (0,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 146,53.

1) **Faustregel ist erfüllt**, da 0% der Zellen eine erwartete Häufigkeit < 5 haben und
2) weil die minimale erwartete Häufigkeit mit 146,53 > 1 ist.
3) *Df ist nicht gleich 1 (nicht in „Kontinuitätskorrektur" ablesen)*

H0: Clusterzugehörigkeit und Altersklassen sind stochastisch unabhängig.
H1: Clusterzugehörigkeit und Altersklassen sind stochastisch abhängig.

H0 wird verworfen, wenn p-Wert < 0,05

p-Wert = < 0,001 < 0,05 → H0 wird verworfen.

➔ K-Means-Clusterzugehörigkeit aus Aufgabe b) und Altersklasse sind stochastisch **abhängig**.

3. Berechnen und interpretieren Sie ein geeignetes Assoziationsmaß für die beiden Variablen *k*-Means-Clusterzugehörigkeit aus Teilaufgabe b) und Altersklasse.

Symmetrische Maße

		Wert	Asymptotischer Standardfehler[a]	Näherungsweises t[b]	Näherungsweise Signifikanz
Ordinal- bzgl. Ordinalmaß	Gamma	-,839	,018	-21,929	<,001
Anzahl der gültigen Fälle		4999			

a. Die Null-Hyphothese wird nicht angenommen.

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

γ = -0,839

Es gibt also einen stark negativen Zusammenhang zwischen der k-Means-Clusterzugehörigkeit und den Altersklassen.
Das bedeutet, dass mit einer höheren Altersklasse eine niedrige Clusternummer einher geht.
→ Also ein höheres Alter geht mit höherem Einkommen, höherem Verhältnis von Schulden/Einkommen und höheren Kreditkarten Schulden einher. Ein niedriges Alter geht mit niedrigerem Einkommen, niedrigerem Verhältnis Schulden/Einkommen und niedrigeren Kreditkarten Schulden einher.

Clusternummer des Falls * Percentile Group of age Kreuztabelle

Anzahl

		Percentile Group of age 1	2	3	Gesamt
Clusternummer des Falls	1	10	73	406	489
	2	1758	1425	1327	4510
Gesamt		1768	1498	1733	4999

Klausur vom 14.07.2020

Aufgabe customer_dbase.sav

a) Führen Sie mit zwei Variablen Ihrer Wahl sowohl eine hierarchische als auch eine k-Means Clusteranalyse durch. Interpretieren Sie die erhaltenen Cluster Ihrer k-Means Clusteranalyse.

1. Wahl von zwei metrisch skalierten Variablen, da eine k-Means-Clusteranalyse nur mit metrisch skalierten Variablen möglich ist.
   a. ed = Years of education
   b. income = Household income in thousands

2. Durchführen der hierarchischen Clusteranalyse:
   Methode „Nächstgelegener Nachbar" mit dem Maß „Euklidische Distanz"

Zusammenfassung der Fallverarbeitung[a,b]

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Prozent
	5000	100,0	0	,0	5000	100,0

a. Euklidisches Distanzmaß verwendet

b. Single Linkage

➔ Fallzahl von N = 5000

Das Dendogramm lässt den größten Sprung am Ende der Tabelle vermuten.

4996	1	3213	67,186	4995	0	4998
4997	1107	2197	78,006	0	0	4999
4998	1	3069	138,004	4996	0	4999
4999	1	1107	215,021	4998	4997	0

Der größte Sprung der Koeffizienten findet vom 4998. auf den 4999. Schritt statt. Damit ergibt sich eine Clusterzahl von 5000-4998 = 2Clustern

3. Ein erneutes durchführen der hierarchischen Clusteranalyse zeigt, dass die beiden Kunden mit der ID 02-SILWTV-4YR und 29-EIXEIO-VD3 gemeinsam ein Cluster ergeben und die anderen 4998 Kunden das andere Cluster.

	custid	ed	income	CLU2_1
1	02-SILWTV-4YR	19	1073,00	2
2	29-EIXEIO-VD3	18	995,00	2
3	64-QJWTRG-NPN	15	31,00	1
4	48-AIPJSP-UVM	17	15,00	1

4. Durchführen der k-Means-Clusteranalyse mit K= 2 Clustern (basierend auf dem Ergebnis der hierachischen Clusteranalyse) und 35 Iterationen.

Anzahl der Fälle in jedem Cluster

Cluster	1	548,000
	2	4452,000
Gültig		5000,000
Fehlend		,000

Clusterzentren der endgültigen Lösung

	Cluster 1	Cluster 2
Years of education	16	14
Household income in thousands	171,82	40,67

Die k-Means-Clusteranalyse ergibt zwei Cluster mit 548 Fällen in Cluster 1 und 4452 Clustern in Cluster 2. Dabei bildet das Cluster 1 das Spitzencluster mit einem durchschnittlich deutlich höheren Einkommen von 171.8200 Geldeinheiten und im Durchschnitt zwei Jahren längerer Ausbildungszeit. Das Cluster 2 umfasst Kunden mit im Durchschnitt niedrigeren Einkommen und kürzeren Ausbildungszeiten (Ablesbar für den jeweiligen Kunden an QCL_1)

	custid	ed	income	CLU2_1	QCL_1
1	8402-SILWTV-4YR	19	1073,00	2	1
2	2329-EIXEIO-VD3	18	995,00	2	1
3	3964-QJWTRG-NPN	15	31,00	1	2
4	0648-AIPJSP-UVM	17	15,00	1	2
5	5195-TLUDJE-HVO	14	35,00	1	2
6	4459-VLPQUH-3OL	16	20,00	1	2
7	8158-SMTQFB-CNO	16	23,00	1	2
8	9662-FUSYIM-1IV	17	107,00	1	1
9	7432-QKQFJJ-K72	14	77,00	1	2

b) Prüfen Sie mit einem Test zum Niveau α= 0,05, ob zwei von Ihnen ausgewählte Variablen stochastisch unabhängig sind.

1. Wahl des Chi-Quadrat-Unabhängigkeitstests und der folgenden ordinalen Variablen:
   a. Edcat = Level of education
   b. Townsize = Size of Hometown

2. Durchführen des Tests zum Niveau α = 0,05
   H_0: Das Ausbildungsniveau der Kunde und die Größe der Heimatstadt sind stochastisch unabhängig.

   ➔ Es wurden N = 4998 Fälle verarbeitet

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)
Chi-Quadrat nach Pearson	27,354[a]	16	0,037710
Likelihood-Quotient	27,422	16	,037
Zusammenhang linear-mit-linear	,118	1	,732
Anzahl der gültigen Fälle	4998		

a. 0 Zellen (0,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 54,61.

3. Überprüfung der Faustregeln:
   a. Df = 16 und damit >1 → erfüllt
   b. 0,0% haben eine erwartete Häufigkeit < 5 und damit weniger als 20% aller Zellen → erfüllt
   c. Minimale erwartete Häufigkeit = 54,61 und damit > 1 → erfüllt

Der p-Wert beträgt 0,037710 und ist damit < 0,05 → Ablehnung der H_0 Hypothese. Das bedeutet, dass die Größe der Heimatstadt und das Ausbildungsniveau voneinander abhängig sind.

c) Berechnen und interpretieren Sie ein geeignetes Assoziationsmaß für die beiden Variablen aus Teilaufgabe b).

1. Wahl des Gamma Koeffizienten, da ordinale Variablen vorliegen.

Symmetrische Maße

		Wert	Asymptotischer Standardfehler[a]	Näherungsweises t[b]	Näherungsweise Signifikanz
Ordinal- bzgl. Ordinalmaß	Gamma	0,005335	,015	,366	,714
Anzahl der gültigen Fälle		4998			

a. Die Null-Hyphothese wird nicht angenommen.

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

Y = 0,005335 → (sehr) schwacher positiver Zusammenhang
Interpretation des Richtungszusammenhangs anhand der Kreuztabelle: Je kleiner die Stadt desto höher der Bildungsgrad.

Size of hometown * Level of education Kreuztabelle

Anzahl

		Level of education					
		Did not complete high school	High school degree	Some college	College degree	Post-undergraduate degree	Gesamt
Size of hometown	> 250,000	277	457	251	345	100	1430
	50,000-249,999	206	323	227	233	66	1055
	10,000-49,999	189	273	178	180	76	896
	2,500-9,999	153	252	187	203	66	861
	< 2,500	128	266	158	151	53	756
Gesamt		953	1571	1001	1112	361	4998

Aufgabe (Klausur vom 09.07.2019)
Öffnen Sie die Datei *CrimeData.sav*.

a) **Ist die mediane Armutsrate (Poverty Rate) in den fünf Kriminalitätskategorien (CrimeType) in etwa gleich groß? Überprüfen Sie dies, indem Sie einen geeigneten Test zum Niveau 0,05 durchführen.**

   1. **Wie heißt der Test?**
   2. **Welche Voraussetzungen gibt es für diesen Test?**
   3. **Wie hoch ist der p-Wert?**
   4. **Interpretieren Sie in knappen Worten die Testentscheidung.**

b) **Klassieren Sie die Werte der Variablen Armutsrate (Poverty Rate) in vier etwa gleich stark besetzte Klassen. Wie viele Fälle liegen in den einzelnen vier Klassen?**

klassierte Armutsrate	
Klasse	**Anzahl Fälle**
1	
2	
3	
4	

   Überprüfen Sie mit einem geeigneten Test zum Niveau 0,05, ob die beiden Variablen „klassierte Armutsrate“ und „Kriminalitätskategorie (CrimeType)“ stochastisch unabhängig sind?

   1. **Wie heißt der Test?**
   2. **Ist die Faustregel erfüllt? (Begründung!)**
   3. **Wie hoch ist der p-Wert?**
   4. **Interpretieren Sie in knappen Worten die Testentscheidung.**
   5. **Wie stark ist der Zusammenhang zwischen den beiden Variablen „klassierte Armutsrate“ und „Kriminalitätskategorie (CrimeType)“?**

c) **Wählen Sie nur die Fälle „ArmedRobbery“ der Variablen Kriminalitätskategorie (CrimeType)“ aus.**

   1. **Führen Sie eine hierarchische Clusteranalyse mit den drei Variablen „POP10 (Population as of 2010 = Bevölkerung)“ und „HU10 (Number of Housing Units = Anzahl Wohneinheiten)“ und „Armutsrate (Poverty Rate)“ durch. Tragen Sie die fehlenden Werte in die folgende Tabelle ein:**

Schritt	Koeffizienten
4 377	
4 378	
4 379	

Wie viele Cluster sind zu bilden? (Begründung!)

2. **Führen Sie eine K-Means Clusteranalyse mit den drei Variablen „POP10 (Population as of 2010 = Bevölkerung)“ und „HU10 (Number of Housing Units = Anzahl Wohneinheiten)“ und „Armutsrate (Poverty Rate)“ durch. Es sollen dabei genau zwei Cluster gebildet werden.**
   Füllen Sie bitte die nachfolgende Tabelle aus:

arithmetisches Mittel

	Cluster	
	1	2
Population as of 2010		
Number of Housing Units		
Poverty Rate		
Anzahl der Fälle		

Interpretieren Sie die zwei Cluster.

Lösung der Klausur vom 09.07.2019

Teilaufgabe a:

Zunächst muss die nominale Zeichenvariable CrimeType in eine nominale numerische Variable umkodiert werden: 1=ArmedRobbery, 2=AutoTheft, 3=MajorTheft, 4=MinorTheft, 5=Violent.

1. Kruskal-Wallis-Test
2. Die Variable „Armutsrate (PovertyRate)“ ist metrisch skaliert. Erlaubt wären hier sowohl ordinal als auch metrisch skalierte Variablen. Die Armutsrate muss in den einzelnen Kriminalitätskategorien (CrimeType) stochastisch unabhängig voneinander sein.
3. Der p-Wert beträgt näherungsweise ungefähr null:

Statistik für Test[a,b]

	Poverty Rate
Kruskal-Wallis H	9527,297
df	4
Asymptotische Signifikanz	,000

a. Kruskal-Wallis-Test

b. Gruppenvariable: CrimeType_num

4. D.h. in mindestens zwei der fünf Kriminalitätskategorien unterscheidet sich die mediane Armutsrate signifikant.

Teilaufgabe b:

Percentile Group of Poverty

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	11078	25,0	25,0	25,0
	2	11043	24,9	24,9	49,9
	3	11032	24,9	24,9	74,8
	4	11145	25,2	25,2	100,0
	Gesamt	44298	100,0	100,0	

1. Chi-Quadrat-Unabhängigkeitstest
2. Freiheitsgrad df = 12, minimale erwartete Häufigkeit = 1090,80 ≥ 1 und keine Zelle hat eine erwartete Häufigkeit kleiner als fünf, erlaubt wären 20%.

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)
Chi-Quadrat nach Pearson	11865,889[a]	12	,000
Likelihood-Quotient	11978,998	12	,000

Anzahl der gültigen Fälle	44298		

a. 0 Zellen (0,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 1090,80.

3. Der p-Wert beträgt etwa null.
4. Die klassierte Armutsrate und die Kriminalitätskategorie hängen voneinander ab.
5. Der Kontingenzkoeffizient beträgt C=0,460; d.h. es gibt einen schwachen Zusammenhang zwischen der klassierten Armutsrate und der Kriminalitätskategorie.

Symmetrische Maße

		Wert	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Kontingenzkoeffizient	,460	,000
Anzahl der gültigen Fälle		44298	

Teilaufgabe c:

Schritt	Koeffizient
4378	2175,365
4379	4263,889

Bei der hierarchischen Clusteranalyse beträgt die Anzahl der Cluster: #Cluster = n -4378 = 4380 - 4378 = 2.

Clusterzentren der endgültigen Lösung

	Cluster	
	1	2
Population as of 2010	8137	4233
Number of Housing Units	2832	1711
Poverty Rate	14,65	15,57

Anzahl der Fälle in jedem Cluster

Cluster	1	1168,000
	2	3212,000
Gültig		4380,000
Fehlend		,000

k-Means-Clusteranalyse:

Cluster 1 = Cluster mit hoher Einwohneranzahl, vielen Gebäuden und niedriger Armutsrate

Cluster 1 = Cluster mit kleiner Einwohneranzahl, wenigen Gebäuden und hoher Armutsrate

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.

Aufgabe: (19.07.2017)
Öffnen Sie die Datei *rfm_transactions.sav*.

a) Prüfen Sie mit einem Test zum Niveau 0,05, ob die Mediane der Variablen „Purchase Amount“ (Ausgaben) in jeder der fünf Produktlinien gleich hoch sind.

1. Wie heißt der Test?
2. Was sind die Voraussetzungen des Tests?
3. Wie hoch ist der p-Wert des Tests?
4. Welche Schlussfolgerung lässt sich aus dem p-Wert ziehen?
5. Wie hoch sind die empirischen Mediane der Variablen „Purchase Amount“ der einzelnen Produktlinien?

Empirische Mediane

	Product-Line				
	A-100	B-200	C-300	D-400	E-500
Purchase Amount					

b) Püfen Sie mit einem Test zum Niveau 0,05, ob Ausgabe-Klassen und Produktlinie stochastisch unabhängig voneinander sind. Unterteilen Sie dazu in der Stichprobe die Variable „Purchase Amount“ (Ausgaben) in drei etwa gleich große Ausgabe-Klassen.

1. Wie heißt der Test?
2. Kontrollieren Sie die Faustregel.
3. Wie hoch ist der p-Wert des Tests?
4. Welche Schlussfolgerung lässt sich aus dem p-Wert ziehen?
5. Berechnen Sie eine Maßzahl für den Zusammenhang zwischen den drei Ausgabe-Klassen und den fünf Produktlinien. Und interpretieren Sie den Wert.

c) Kodieren Sie um wie folgt:

- „Product-Line“ in eine binäre Variable mit 1=E-500, 0≠E-500
- „Purchase Amount“ in eine binäre Variable mit 1=amount $>$ 80, 0=amount $\leq$ 80

Führen Sie eine hierarchische Clusteranalyse mit diesen beiden binären Variablen durch.

1. Wie viele Cluster sind zu bilden? (Begründung!)
2. Wie viele Fälle liegen in den jeweiligen Clustern?
3. Gibt es einen Unterschied zwischen den gebildeten Clustern und den Feldern einer Kreuztabelle zwischen den beiden binären Variablen?

QM Master Klausur vom 19.07.2017

a.1) Kruskal-Wallis Test

a.2) ordinale oder metrische Skalierung, stochastische Unabhängigkeit

a.3)

Statistik für Test[a,b]

	Purchase Amount
Kruskal-Wallis H	4443,742
Df	4
Asymptotische Signifikanz	,000

a. Kruskal-Wallis-Test

b. Gruppenvariable: ProductLine_num

a.4) Die medianen Ausgaben sind in mindestens zwei der fünf Produkt-Linien signifikant unterschiedlich

a.5)

Median

Product Line	Purchase Amount
A-100	29,00
B-200	54,00
C-300	83,00
D-400	123,00
E-500	185,00
Insgesamt	81,00

b.1) Chi-Quadrat Unabhängigkeitstest

b.2)

Product Line * Percentile Group of Amount Kreuztabelle

Anzahl

		Percentile Group of Amount			
		1	2	3	Gesamt
Product Line	A-100	1026	0	0	1026
	B-200	520	467	0	987
	C-300	89	821	41	951
	D-400	0	339	604	943
	E-500	1	0	998	999
Gesamt		1636	1627	1643	4906

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)
Chi-Quadrat nach Pearson	6325,744[a]	8	,000
Likelihood-Quotient	7245,651	8	,000
Anzahl der gültigen Fälle	4906		

a. 0 Zellen (0,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 312,73.

b.2) Faustregel ist erfüllt, da keine erwartete Häufigkeit kleiner als fünf ist, hier wären 20% erlaubt. Und weil die minimale erwartete Häufigkeit mit 312,73 größer als eins ist.

b.3) p-Wert = 0,000

b.4) Ausgabeklasse und Produkt-Linie sind stochastisch abhängig

b.5)

Symmetrische Maße

		Wert	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Kontingenzkoeffizient	,750	,000
Anzahl der gültigen Fälle		4906	

Der Kontingenzkoeffizient ist mit 0,750 lässt auf einen fast starken Zusammenhang zwischen Ausgabenklasse und Produktlinie schließen.

c) Hierarchische Clusteranalyse

Größter Sprung der Koeffizienten von 0 im Schritt 4902 auf 1 im Schritt 4903.

Anzahl Cluster = 4906 – 4902 = 4 Cluster

Single Linkage

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	2442	49,8	49,8	49,8
	2	1465	29,9	29,9	79,6
	3	998	20,3	20,3	100,0
	4	1	,0	,0	100,0
	Gesamt	4906	100,0	100,0	

Kontingenztabelle / Kreuztabelle

Amount_binär * PLinie_Binär Kreuztabelle

Anzahl

		PLinie_Binär		
		A-100, B_200, C-300, D-400	E-500	Gesamt
Amount_binär	kleinergleich 80	2442	1	2443
	größer 80	1465	998	2463
Gesamt		3907	999	4906

Die vier Cluster entsprechen genau den vier Zellen der Kreuztabelle.

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!)

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	

Kruskal-Wallis-Test

Die Variable muss ordinal oder metrisch skaliert sein. Dies ist hier der Fall, da die Variable „Gehalt“ metrisch skaliert ist. Der p-Wert des Kruskal-Wallis-Test beträgt in etwa null. D.h. in mindestens zwei der drei Cluster ist das mediane Gehalt signifikant unterschiedlich hoch.

Statistik für Test[a,b]

	Gehalt
Kruskal-Wallis H	154,441
df	2
Asymptotische Signifikanz	,000

a. Kruskal-Wallis-Test

b. Gruppenvariable: Clusternummer des Falls

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.

Lösung der Klausur 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".

Statistik für Test[a,b]

	Income category (thousands)
Kruskal-Wallis H	1505,887
df	2
Asymptotische Signifikanz	,000

a. Kruskal-Wallis-Test

b. Gruppenvariable: Clusternummer des Falls

Der p-Wert des Kruskal-Wallis-Test beträgt näherungsweise null; d.h. in mindestens zwei Cluster ist das mediane Gehalt signifikant unterschiedlich.

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.

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!)

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=0,003 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

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?

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.

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

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.

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.

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.

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.