Prof. Dr. Reinhold Kosfeld
Fachbereich Wirtschaftswissenschaften
Universität Kassel

Lösungen zu den Aufgaben zur Multivariaten Statistik
Teil 1: Aufgaben zur Einleitung

1. Was versteht man unter einer univariaten, bivariaten und multivariaten Analyse?

   Lösung:

   In der univariaten Analyse wird ein Merkmal separat ausgewertet z.B. in Form einer eindimensionalen Häufigkeitsverteilung und der Berechnung eines Mittelwertes und/oder eines Streuungsmaßes.

   Eine bivariate Analyse hat die Auswertung verbundener Beobachtungspaare zweier Merkmale zum Gegenstand etwa durch die Darstellung ihrer gemeinsamen Häufigkeitsverteilung und der Berechnung eines Korrelationskoeffizienten oder der Schätzung einer Regressionsgeraden.

   Von einer multivariaten Analyse spricht man, wenn gleichzeitig mehr als zwei Merkmale ausgewertet werden, was typischerweise bei Anwendung der Faktoren-, Varianz-, Cluster- und Diskriminanzanalyse der Fall ist.

2. Nennen Sie typische univariate und bivariate Auswertungsmethoden!

   Lösung:

   Univariate Auswertungsmethoden:
   - Eindimensonale Häufigkeitsverteilungen,
   - Mittelwerte,
   - Streuungsmaße,
   - Gestaltmaße (Schief- und Wölbungsmaße),
   - Konzentrationsmaße

   Bivariate Auswertungsmethoden:
   - Zweidimensionale Häufigkeitsverteilungen,
   - Kovarianz,
   - Korrelationskoeffizienten,
   - Preisindizes,
   - Mengenindizes

3. Nennen Sie multivariate Verfahren der Dependenz- und Interdependenzanalyse! Worin bestehen die Unterschiede?

   Lösung:

   Siehe Folien und Lehrbuch zur „Multivariaten Statistik“.

4. Erläutern Sie den Unterschied des Begriffs „Faktor“ in der Faktoren- und Varianzanalyse!

   Lösung:

   In der Faktorenanalyse bezieht sich der Begriff „Faktor“ auf eine latente, d.h. nicht direkt beobachtbare Variable, die aus der Korrelationsmatrix der manifesten, d.h. beobachtbaren Variablen „extrahiert“ wird.

   In der Varianzanalyse werden dagegen die unabhabhängigen nominalskalierten, d.h. qualitativen Variablen, als „Faktoren“ bezeichnet, deren Einfluss auf eine metrisch skalierte, d.h. quantitative Variable, untersucht werden soll. Die in der Varianzanalyse definierten Faktoren werden als beobachtbar vorausgesetzt.

5. Wodurch unterscheidet sich die Clusteranalyse von der Diskriminanzanalyse?

   Lösung:

   Die Clusteranalyse ist ein strukturentdeckendes Verfahren, dass Beziehungszusammenhänge zwischen Objekten aufzudecken sucht (exploratives Verfahren). Speziell zielt die Clusteranalyse darauf ab, Gruppen von Objekten in einem m-dimensionalen Merkmalsraum „zu entdecken“. Die Clusterzahl sowie die Clusterzugehörigkeit sind bei der Anwendung hierarchischer Klassifikationsverfahren typischerweise unbekannt. Bei der Anwendung von Partitionsverfahren wird zumindest eine Vorstellung von der möglichen Clusterzahl vorausgesetzt.

   Dagegen gehört die Diskriminanzanalyse zu den strukturprüfenden Verfahren (konfirmatorisches, d.h. strukturtestendes Verfahren). Sie untersucht, ob sich bereits bestehende Gruppen von Untersuchungseinheiten durch sog. Merkmalsvariablen trennen lassen (Diskrimination). Clusterzahl und Clusterzugehörigkeit werden hierbei als bekannt vorausgesetzt. Neue Objekte mit unbekannter Gruppenzugehörigkeit lassen nach einer Verdichtung der Merkmalsvariablen zu einer Diskriminanzfunktion den bestehenden Gruppen zuordnen (Klassifikation).

6. Gegeben sind die Daten der Arbeitsproduktivität (X_1) und der Löhne (X_2) in fünf Regionen:

Region	A	B	C	D	E
Arbeitsproduktivität	3,8	4,2	2,8	3,2	3,0
Löhne	20,2	24,6	21,6	23,2	22,4

   a) Berechnen Sie den Korrelationskoeffizienten und interpretieren Sie ihn!

   Lösung:

   Formel des Korrelationskoeffizienten:

$$r_{12} = \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)\cdot(x_{i2}-\bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2 \cdot \sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2}}$$

Arbeitstabelle:

i	x_{i1}	x_{i2}	$(x_{i1}-\bar{x}_1)^2$	$(x_{i2}-\bar{x}_2)^2$	$(x_{i1}-\bar{x}_1)\cdot(x_{i2}-\bar{x}_2)$
1	3,8	20,2	0,16	4,84	-0,88
2	4,2	24,6	0,64	4,84	1,76
3	2,8	21,6	0,36	0,64	0,48
4	3,2	23,2	0,04	0,64	-0,16
5	3,0	22,4	0,16	0	0
Σ	17	112	1,36	10,96	1,2

Arithm Mittel: $\bar{x}_1 = 17{,}4/5 = 3{,}4$ und $\bar{x}_2 = 112/5 = 22{,}4$

Korrelationskoeffizoent: $r_{12} = \frac{1{,}2}{\sqrt{1{,}36 \cdot 10{,}96}} = \frac{1{,}2}{3{,}860777} = 0{,}311$

<u>Interpretation</u>: Der Korrelationskoeffizient zwischen der Arbeitsproduktivität und den Löhnen liegt an der Grenze eines schwachen und mittelstarken positiven Zusammenhangs.

b) Testen Sie den Korrelationskoeffizienten auf Signifikanz (α=0,05)!

1. Schritt: Null- und Alternativhypothese

H_0: $\rho_{jk} = 0$
H_1: $\rho_{jk} \neq 0$ (zweis. Test)

2. Schritt: Festlegung des Signifikanzniveaus

$\alpha = 0{,}05$

3. Schritt: Berechnen der Teststatistik

Prüfgröße: $T_0 = \frac{r_{jk}}{\sqrt{1-r_{jk}^2}} \cdot \sqrt{n-2} \sim t_{n-2}$

mit n=5 und r_{12}=0,311 (s. Teil a)): $t_0 = \frac{0{,}311}{\sqrt{1-0{,}311^2}} \cdot \sqrt{3} = 0{,}5668$

4. Schritt: Kritischer Wert (n=5 und α = 0,05)

$t_{n-2;1-\alpha/2} = t_{3;0,975} = 3{,}182$

5. Schritt: Testentscheidung

$(t_0 = 0{,}5668) < (t_{3;0,975} = 3{,}182) \quad \Rightarrow \quad H_0$ annehmen

7. Nach welchem Ordnungsprinzip ist ein Datensatz in einem Dateneditor eines Statistikprogramms wie z.B. R oder SPSS typischerweise angeordnet?

   Lösung:

   Ein Datensatz ist in einem Dateneditor eines Statistikprogramms wie z.B. SPSS nach einer Matrixstruktur organisiert. In einer Zeile des Datensatzes stehen die beobachteten Werte einer Untersuchungseinheit (statistische Einheit, Objekt, Fall) für alle Variablen (Merkmale). Eine Spalte des Datensatzes gibt dagegen alle bei den statistischen Einheiten beobachteten Werte eines bestimmten Merkmals wieder.

8. Welche Informationen über die Variablen werden typischerweise in einem Dateneditor festgehalten?

   Lösung:

   Die typischerweise in einem Dateneditor über die Variablen festgehaltenen Informationen beziehen sich im Wesentlichen auf den Variablennamen, den Variablentyp, das Messniveau, Variablenlabels, Wertelabels und die Definition fehlender Werte. Darüber hinaus werden technische Angaben etwa zur Spaltenzahl, zum Spaltenformat und zur Anzahl der Dezimalstellen erfasst.