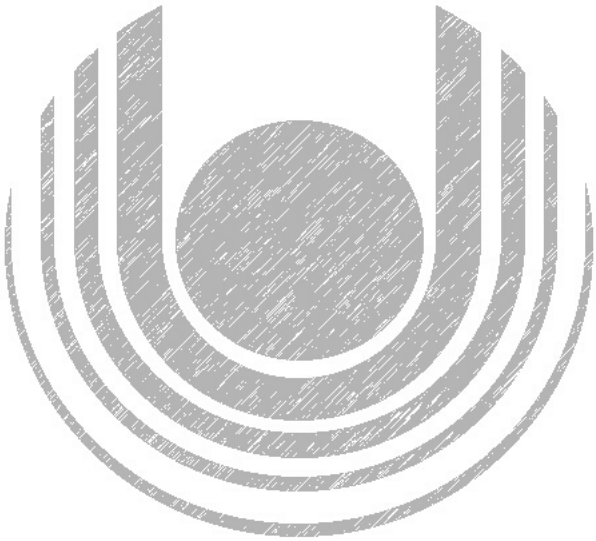

Name, Vorname

Matrikelnummer



Modulklausur 31821 – Multivariate Verfahren

Datum	
Punkte	
Note	

Termin: 20.09.2019, 11:30 - 13:30
Prüfer: Univ.-Prof. Dr. H. Singer

Hinweise zur Bearbeitung der Modulklausur 31821

1. Füllen Sie zunächst den **Kopf des Deckblatts** aus!

2. Es können insgesamt 100 Punkte erreicht werden. Bei Erreichen von 50 Punkten ist die Klausur bestanden. **Bitte kontrollieren Sie sofort, ob Sie ein vollständiges Klausurexemplar erhalten haben**.

3. Zugelassen ist Kurseinheit 1 des Moduls 31821 (Kursnr. 00883) mit farblichen Markierungen, kleinen Aufklebern und/oder textbezogenen Anmerkungen. Nicht zugelassen sind selbst ausgedruckte oder kopierte Kursmaterialien.
   Die Verwendung eines Taschenrechners ist dann und nur dann erlaubt, wenn dieser einer der drei folgenden Modellreihen angehört:

   - Casio fx86 oder Casio fx87
   - Texas Instruments TI 30 X II
   - Sharp EL 531

   Die Verwendung anderer Taschenrechnermodelle wird als Täuschungsversuch gewertet und mit der Note „nicht ausreichend“ $(5,0)$ sanktioniert.
   Ob ein Taschenrechner einer der drei Modellreihen angehört, können Sie selbst überprüfen, indem Sie die vom Hersteller auf dem Rechner angebrachte Modellbezeichnung mit den oben angegebenen Bezeichnungen vergleichen: Bei vollständiger Übereinstimmung ist das Modell erlaubt. Ist die auf dem Rechner angebrachte Modellbezeichnung umfangreicher, enthält aber eine der oben angegebenen Bezeichnungen vollständig, ist das Modell ebenfalls erlaubt. In allen anderen Fällen ist das Modell nicht erlaubt.

4. Bitte benutzen Sie für Ihre Rechnungen nur die beigefügten Lösungsbögen.

5. Wenn Sie die einzelnen Blätter der Klausur voneinander trennen, **vermerken Sie auf jedem Blatt Ihre Matrikelnummer**. Legen Sie bitte am Ende der Klausur die Blätter wieder zusammen.

Wir wünschen Ihnen viel Erfolg!

Aufgabe 1 (20 Punkte)

Eine bivariat normalverteilte Zufallsvariable $\mathbf{x} = [X, Y]' \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ist durch den Erwartungswert $\boldsymbol{\mu} = [3, 4]'$ und die Kovarianzmatrix $\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 0.25 \\ 0.25 & 1 \end{bmatrix}$ gegeben.

a) Wie lautet die Randdichte $\phi(x)$ der Zufallsvariablen X? Geben Sie die Dichefunktion explizit an. (5 P.)

b) Eine neue Zufallsvariable Z ist durch die lineare Transformation $Z = a + \mathbf{bx}$ mit $a = 3$ und $\mathbf{b} = [4, 1]$ gegeben. Wie lautet die Verteilung dieser Variable? Geben Sie den Erwartungswert und die Varianz von Z explizit an. (5 P.)

c) Prognostizieren Sie die Variable Y durch die Variable X. Welche funktionale Form hat die optimale Prognosefunktion $\hat{Y} = g(X)$? (5 P.)

d) Berechnen Sie die Korrelationsmatrix von $\mathbf{x} = [X, Y]'$. (5 P.)

Aufgabe 2 (20 Punkte)

a) Definieren Sie die L_q-Distanz und beschreiben Sie die Complete-Linkage-Methode. Nach welchem Kriterium werden die Klassen miteinander fusioniert? (5 P.)

b) Gegeben seien die Objekte A, B, C und D, die jeweils durch vier Merkmale beschrieben werden. Die entsprechende Datenmatrix $\mathbf{X}$ hat dabei folgende Gestalt:

$$\mathbf{X} = \begin{bmatrix} 10 & 2 & 3 & 1 \\ 2 & 1 & 2 & 5 \\ 5 & 1 & 1 & 7 \\ 4 & 1 & 1 & 8 \end{bmatrix}$$

Hinweis : Zeile 1 enthält die Merkmale von Objekt A, Zeile 2 die von Objekt B usw.

Konstruieren Sie agglomerativ eine Hierarchie. Verwenden Sie dabei die Complete-Linkage-Methode, indem Sie zunächst die euklidische Distanzmatrix bilden. Geben Sie stets die Klassen und den Indexwert der Fusion, sowie die Distanzmatrix an. Zeichnen Sie das entsprechende Dendrogramm.

Hinweis : Für zwei positive Zahlen x und y mit $x < y$ gilt $\sqrt{x} < \sqrt{y}$, d.h Sie müssen die Wurzel nicht explizit ausrechnen.

(15 P.)

Aufgabe 3 (20 Punkte)

a) Definieren Sie das Grundmodell der einfaktoriellen Varianzanalyse mit fixen Effekten. Geben Sie zusätzlich die Formeln für die totale/erklärte/residuale Streuung an. Wie lautet der Determinationskoeffizient? (5 P.)

b) Ein Unternehmen stellt Socken her und möchte überprüfen, ob die Anwesenheit von Haustieren am Arbeitsplatz die Produktion der Mitarbeiter fördert. Dazu wurde die Leistung von vier Mitarbeitern mit und ohne Haustier auf die Variable „Anzahl hergestellter Socken "gemessen:

Person	ohne Haustier	mit Haustier
A	17	37
B	13	14
C	31	25
D	59	69

Stellen Sie die ANOVA-Tabelle auf und prüfen Sie mit dem F-Test die Gleichheit der Mittelwerte zum 5%-Signifikanzniveau.

Hinweis: Das f-Quantil lautet $f(0.95; 1, 6) = 5.99$. (15 P.)

Aufgabe 4 (40 Punkte)

Herleitung und Anwendung des Logit-Modells:

a) Geben Sie die Bayes-Formel und den Satz von der totalen Wahrscheinlichkeit an. Beschreiben Sie anschließend die logistische Funktion (mit Formel)! Wie sieht das Logit-Modell aus? (5 P.)

b) Es sei Y eine abhängige Variable mit Ausprägung $y = 0, 1$ und x die dazugehörige Regressionsvariable. Weiter sei $p(y|x) := P(Y = y|X = x)$. Beweisen Sie: Für den Fall, dass die bedingte Normalverteilung vorliegt, d.h. $p(x|y = 0) = \phi(x; \mu_0, \sigma^2) = \phi_0$ und $p(x|y = 1) = \phi(x; \mu_1, \sigma^2) = \phi_1$, lässt sich die logistische Funktion aus der Bayes-Formel herleiten.

Hinweis: Nutzen Sie die Bayes-Formel aus Teil a). Schreiben Sie dann die Funktionen ϕ_0 und ϕ_1 explizit aus und setzen diese ein. (10 P.)

c) In diesem Aufgabenteil soll mit dem logistischen Modell die Einflussfaktoren für die Diagnose Krebs bei Personen unter 40 Jahren untersucht werden. Die abhängige Variable `Krebs` nimmt den Wert 1 bei „Krank“und den Wert 0 bei „Gesund“an. Als erklärende Variablen stehen zur Verfügung: `Raucher` (=1, wenn die Person mindestens 5 Jahre Raucher ist, und 0 andernfalls) und `gesunde Ernährung` (=1, wenn sich die Person gesund ernährt und 0 sonst). In der folgenden Tabelle sind die zu analysierenden Daten zu finden.

Person	Raucher	gesunde Ernährung	Krebs
A	1	1	0
B	0	1	0
C	1	1	1
D	0	1	0
E	1	0	1

1. Berücksichtigen Sie in den folgenden Aufgabenteilen nur die unabhängige Variable `gesunde Ernährung` in Dummy-Kodierung und die abhängige Variable `Krebs`. Bestimmen Sie nun die Likelihoodfunktion und die Log-Likelihood für ein Modell mit Konstante! (10 P.)

2. Nehmen Sie an, der Optimierungsalgorithmus liefere folgende Parameterschätzung in Dummy-Kodierung:

   $$\hat{\boldsymbol{\beta}} = [\hat{\beta}_0,\ \hat{\beta}_1] = [0.5,\ -0.5]'$$

   Wie hoch ist die Wahrscheinlichkeit, dass eine rauchende Person Krebs hat? (7 P.)

3. Sei $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0,\ \hat{\beta}_1] = [0.5,\ -0.5]'$ wie in Teil b) beschrieben. Angenommen die zweite Ableitung nach beiden Variablen ist $\mathbf{H} = \begin{bmatrix} -1 & -0.5 \\ -0.5 & -0.3 \end{bmatrix}$. Testen Sie mit der Wald-Statistik zum 5%-Niveau, ob die tatsächlichen Werte $\boldsymbol{\xi} = \boldsymbol{\beta} = [1,\ 0]'$ sind. Dabei ist $\chi^2(2) = 5.991$ (8 P.)