

Rechenaufgaben zum Erwerb des Scheins „Einführung in die statistischen Programmpakete (Einführungskurs in SPSS)“.

Beachten Sie bitte folgende Hinweise:

- Abgabe bis spätestens 17. August 2007
- Maximal 2 Personen können das Blatt gemeinsam bearbeiten.
- Die kommentierte Syntax ist als gesonderte Datei abzugeben (per email).
- Die Lösungen und Ergebnisse der Aufgaben müssen als Anhang per email an *christian.heumann@stat.uni-muenchen.de* geschickt werden. Die Dateinamen dieses Anhangs sollen immer den oder die Nachnamen der Personen enthalten.
- Die Syntax muss ohne Änderungen lauffähig sein.
- Zusätzlich sind die Aufgaben in Papierform in gebundener Form abzugeben. Es kann sich dabei auch um die ausgedruckte Form des Anhangs handeln, allerdings soll der Ausdruck schon den Charakter eines Berichts haben (und nicht nur eine Loseblattsammlung)!

1. Waldschadensdaten.

Auf der Webseite

<http://www.statistik.lmu.de/service/datenarchiv/baum/baum.html>

befindet sich die Beschreibung des Datensatzes „Standorteinflüsse auf das Ausmaß von Baumschäden“. Es handelt sich um sog. Cluster-Daten: jeweils 8 Bäume an einem Standort wurden bezüglich ihrer Baumschäden beurteilt. Die zu einem Cluster gehörenden Bäume sind an der gleichen Clusternummer (Variable *clust*) erkennbar. Während die Waldschadensstufen für jeden Baum im Cluster verschieden sein *können*, handelt es sich bei den Standortvariablen um sog. Cluster-Variablen, d.h. die Standortvariablen haben für alle Bäume in einem Cluster die gleichen Ausprägungen. Da die Bäume in einem Cluster nicht als unabhängig angesehen werden können, soll zunächst ein neuer Datensatz erzeugt werden, der folgende Variablen enthält:

- *Schadensscore*: In jedem Cluster wird das arithmetische Mittel (die neue Variable) *score* aus der Waldschadensvariable *schaß* gebildet.
- Dazu werden die Cluster-Standortvariablen *schir*, *ort*, *expo* und *hoehe* bestimmt.

D.h. der neue Datensatz hat statt $(771 \cdot 8 = 6168)$ Zeilen nur noch 771 Zeilen und 5 Spalten.

- (a) Erstellen Sie geeignete deskriptive Statistiken für alle Variablen.
- (b) Stellen Sie die univariaten Verteilungen in geeigneter Weise grafisch dar.
- (c) Stellen Sie die Verteilung der Variable *score* für die einzelnen Ausprägungen der kategorialen Variablen so dar, dass ein visueller Vergleich möglich ist.
- (d) Stellen Sie die Variablen *score* und *hoehe* gemeinsam geeignet grafisch dar.

2. Mietspiegeldaten.

Die Daten finden Sie auf der Seite des SPSS-Kurses.

- (a) Erstellen Sie geeignete deskriptive Statistiken für alle Variablen.
- (b) Stellen Sie die Verteilungen der metrischen Variablen in einer geeigneten Grafik dar. Welche Gestalt (symmetrisch, linksschief, rechtsschief) weisen diese auf? Betrachten Sie die Variablen *nm* und *wfl*. Führen Sie eine Transformation dieser Variablen mit Hilfe des natürlichen Logarithmus und der Wurzelfunktion durch (erzeugen Sie neue Variablen). Zeichnen Sie Q-Q-Plots (Normalverteilung) für die ursprünglichen und die neuen Variablen. Führen Sie einen Kolmogorov-Smirnov Anpassungstest auf Normalverteilung für die ursprünglichen und die neuen Variablen durch. Beurteilen Sie die Ergebnisse.
- (c) Erzeugen Sie ein (sinnvolles) Stuediagramm für *wfl* und *nm* mit Regressionsgerade. Beurteilen Sie die Anpassung! Welche Modellannahmen der einfachen linearen Regression sind vermutlich verletzt? Überprüfen Sie, ob die Anpassung durch irgendeine Kombination der in (b) berechneten Variablen verbessert werden kann.
- (d) Erstellen Sie einen gruppierten Boxplot für *nmqm* nach *rooms*.
- (e) Überprüfen Sie mit einem geeigneten parametrischen und nichtparametrischen Test, ob sich die mittlere Nettomiete pro Quadratmeter (*nmqm*) der Wohnungen mit einem Zimmer von der mittleren Nettomiete pro Quadratmeter der übrigen Wohnungen unterscheidet. Vergleichen Sie analog Wohnungen, die bis einschliesslich 1960 gebaut wurden mit den Wohnungen, die nach 1960 gebaut wurden.
- (f) Teilen Sie die Daten zur Analyse mittels der Variable *rooms* auf. Führen Sie die Regression(en) von *nm* auf *wfl* durch. Vergleichen Sie die Ergebnisse mit der entsprechenden Regression der Gesamtdaten in (c).
- (g) Bilden Sie eine neue kategoriale Variable *nmqmkat* mit 4 Kategorien aus der Variable *nmqm*, so dass die Kategorien etwa gleiche Besetzungszahlen aufweisen. Besteht ein Zusammenhang zwischen dieser Variable und den anderen kategorialen Variablen (z.B. *wohngut*)? Die Ausgabe soll dabei jeweils die bedingten Verteilungen der Variable *nmqmkat* gegeben die Kategorien der anderen kategorialen Variable enthalten. Interpretieren Sie die Ergebnisse kurz.
- (h) Jetzt entscheiden Sie! Führen Sie eine weitere Analyse durch, die Ihrer Meinung nach besonders interessant ist.