1. Übungsblatt *506.715 Angewandte Statistik, WS 2009/2010*

Univ.-Prof. DI Dr. Ernst Stadlober

1.) EDA und CDA, Einstichprobenproblem, `aimu_1985.dat;[R 2.9, SPSS 17.0]`

(a) Lesen Sie die Textdatei `aimu_1985.dat` in `R` oder `SPSS` ein. Man definiere neue kategorische Variable `jung_alt (1,2)` mit `alter` 16-30, 31-56; `al_kl (1,2,3,4,5)` mit `alter` 16–19, 20–25, 26–32, 33–40, 41–56; `gr_kl (1,2,3,4)` mit `gr_cm` 160–172, 173–176, 177–181, 182–195.

(b) Analysieren Sie **zwei** der Variablen `alter, gr_cm, ge_kg, fvc, fev1` mit den Methoden der explorativen Datenanalyse. Benutzen Sie *Histogramme*, *Stem–and–Leaf–Plots, (Fehlerbalken (Error Bars) in* `SPSS`*), Boxplots, empirische Verteilungsfunktionen und Q–Q–Plots.*

(c) Berechnen Sie für *zwei* der Variablen `gr_cm, ge_kg, fvc, fev1` statistische Kenngrößen, in `R` über den Befehl `summary()`, in `SPSS` gemäß Bsp. 2.1. Berechnen Sie in `R` auch s_q, s_{MAD}, g_1^q, g_2^q und cv.

(d) Führen Sie für **zwei** der Variablen `gr_cm, ge_kg, fvc, fev1` (i) den Kolmogorov-Smirnov-Test und (ii) den Shapiro-Wilk-Test auf Normalverteilung durch. (In `SPSS` unter dem Menü *Analysieren* $\longrightarrow$ *Explorative Datenanalyse.*)

(e) Was liefert der t–Test bzgl. der Hypothesen $\mu_{gr} = 176$, $\mu_{ge} = 82$, $\mu_{fvc} = 5.5$ und $\mu_{fev1} = 4.4$?

(f) Fassen Sie Ihre Ergebnisse und Interpretationen in Form eines pdf–Dokuments mit max. 4 Seiten) zusammen.

2.) EDA und CDA, Zweistichprobenproblem, Merkmale `fvc, fev1` mit Kategorien `jung_alt, region; [R 2.9, SPSS 17.0]`.

(a) Geben Sie *Histogramme, (in* `R` *gekerbte) Box-Plots, (Fehlerbalken in* `SPSS`*), Stem–and–Leaf-Plots* bzgl. der beiden Merkmale `fvc, fev1` in Abhängigkeit von `jung_alt` bzw. `region` an.

(b) Was liefern die Q-Q-Plots mit Normalverteilung und die univariaten Tests auf Normalverteilung (K–S–Test und Shapiro–Wilk–Test) für die einzelnen Gruppen?

(c) Geben Sie die Schätzer der Standardabweichungen für die Mediane $\tilde{x}$ und $\tilde{y}$ an und ermitteln Sie daraus die Bereiche der *gekerbten (notched)* Boxplots mit $\tilde{x} \pm 1.7\,\hat{\sigma}(\tilde{x})$. Wie lauten die 95%–Konfidenzintervalle für die Differenzen $m_D = m_X - m_Y$ unter der Annahme $\sigma(\tilde{X}) = \sigma(\tilde{Y})$?

(d) Führen Sie für beide Merkmale `fvc, fev1` bzgl. `jung_alt` bzw. `region` die entsprechenden t–Tests durch und berechnen Sie 99%–Konfidenzintervalle für $\mu_D = \mu_X - \mu_Y$. Als Test auf Gleichheit der Varianzen wird in `SPSS` der Levene–Test benutzt und in `R` der Fligner–Test. Was liefert dazu der klassische F–Test in `R`?

(e) Welche Ergebnisse erhält man mit (i) dem Mann–Whitney–U–Test und (ii) dem Kolmogorov–Smirnov–Test bzgl. des Vergleichs der klassifizierten Stichproben?

(f) Fassen Sie Ihre Ergebnisse und Interpretationen in Form eines pdf–Dokuments mit max. 4 Seiten) zusammen.

3.) [T] Verteilungen und Kenngrößen von Verteilungen.

(a) Sei $X \sim F$ mit $E(X) = \mu$, $Var(X) = \sigma^2$. Man zeige, dass die Schiefe $\gamma_1(X)$ und die Kurtosis $\gamma_2(X)$ invariant sind unter der Standardisierung $Z = (X - \mu)/\sigma$; d.h. $\gamma_i(X) = \gamma_i(Z)$, $i = 1, 2$.

(b) Wegen (a) kann man o.B.d.A. $E(X) = 0$ und $Var(X) = 1$ annehmen. Man zeige die Ungleichung $\gamma_2(X) \geqslant \gamma_1^2(X) - 2$.

Hinweis: Integrieren Sie $\int (x^2 - \gamma_1 x - 1)^2 dF(x)$

(c) Man zeige folgende Identität für $k < n$, $0 < p < 1$:

$$\sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} = \binom{n}{k} (n-k) \int_p^1 x^k (1-x)^{n-k-1} dx$$

(i) durch partielle Integration, (ii) indem beide Seiten als Funktion von p aufgefasst werden und bzgl. p differenziert wird.

(d) Sei $X_i \overset{iid}{\sim} F$ stetige Zufallsvariable mit Dichte $f = F'$. $k = \lfloor np \rfloor + 1$, $f(x_p) > 0$, wobei gilt $F(x_p) = p$. Es gilt $X_{(k)} = F^{-1}(U_{(k)})$ mit $U_{(k)} \sim beta(k, n-k+1)$, $E(U_{(k)}) = \frac{k}{n+1}$, $Var(U_{(k)}) = \frac{k}{(n+1)(n+2)} \left(1 - \frac{k}{n+1}\right)$. Ist F^{-1} zweimal differenzierbar, dann verifiziere man (vergleiche **Satz 2.3.2.**):

$$\begin{aligned} E(X_{(k)}) &\approx x_p + \frac{p(1-p)}{2n\, f^2(x_p)} \left(- \frac{d\, f(F^{-1}(u))}{du} \right) \Bigg|_{u=p} \\ Var(X_{(k)}) &\approx \frac{p(1-p)}{n\, f^2(x_p)} . \end{aligned}$$

(e) Man zeige mit Hilfe von (c) und (d), dass

$$P(X_{(k)} < x_p < X_{(l)}) = \sum_{i=k}^{l-1} \binom{n}{i} p^i (1-p)^{n-i} .$$

4.) EDA, k–Stichprobenproblem, Merkmale `fvc,fev1` mit Kategorien `jung_alt, gr_kl;` `[R 2.9, SPSS 17.0]`.

(a) Man erzeuge Box–Plot–Serien und (Fehlerbalken in `SPSS`) für die Merkmale `fvc` `(fev1)` bzgl. der Kategorien `jung_alt` und `gr_kl` getrennt. (4 Serien; `SPSS:` *Optionen: Einfach, Auswertung über Kategorien einer Variablen*).

(b) Man erzeuge Box-Plot-Serien und (Fehlerbalken in `SPSS`) für die Merkmale `fvc` und `fev1` gemeinsam, aber getrennt nach den Kategorien `jung_alt` und `gr_kl`. (2 Serien; `SPSS:` *Optionen: Gruppiert, Auswertung für verschiedene Variablen*).

(c) Man erzeuge Box-Plot-Serien und (Fehlerbalken in `SPSS`) für `fvc` und `fev1` getrennt, aber gemeinsam nach der Kategorie `jung_alt` und Gruppe `gr_kl`. (2 Serien; `SPSS:` *Optionen: Gruppiert, Auswertung über Kategorien einer Variablen*).

(d) Versuchen Sie aus (a) – (c) entsprechende Schlüsse zu ziehen.

(e) Man generiere Streudiagramme (Scatterplots) von `fvc (fev1)` gegen `gr_cm` bzw. `alter` und lege Regressionsfunktionen durch. (`SPSS`: Wählen Sie nacheinander eine lineare und quadratische Regression, sowie die nichtparametrische Glättung `lowess` aus).

(f) Man erzeuge die Scatterplotmatrix für die Variablen `alter, gr_cm, ge_kg, fvc, fev1` und lege entsprechende Regressionsfunktionen durch.

(g) Fassen Sie Ihre Ergebnisse und Interpretationen in Form eines pdf–Dokuments mit max. 4 Seiten) zusammen.

5.) Verbundene Stichproben (Matched pairs); `[R 2.9, SPSS 17.0]`.
Um den Einfluss einer Yoga–Übung auf den Blutdruck zu bestimmen, wurden an $n = 14$ Personen Blutdruckmessungen in mmHg (systolisch/diastolisch) **vor** und **nach** der Übung gemessen. Die gemessenen Daten sind in der folgenden Tabelle angegeben.

Yoga–Daten von Feuerabendt/Hammer (1987)

Nr.	Geschlecht	Alter	Blutdruck	
			vorher	nachher
1	w	43	140/90	110/70
2	w	39	100/80	120/70
3	m	36	120/70	130/70
4	m	76	130/100	190/130
5	w	40	150/80	130/90
6	w	49	115/75	120/80
7	m	41	100/80	130/60
8	w	27	140/80	120/70
9	m	37	105/80	120/60
10	w	21	105/80	110/70
11	m	38	130/75	120/65
12	w	52	120/90	110/85
13	w	69	145/80	130/80
14	m	32	115/85	125/65

(a) Definieren Sie einen entsprechenden `R`–File `yoga.dat` oder `SPSS`–File `yoga.sav` mit Variablen, deren Labels etc. Definieren Sie die Variable `d_syst` als Differenz des *systolischen Blutdrucks vorher* mit dem *systolischen Blutdruck nachher*, analog die Variable `d_diast`.

(b) Führen Sie eine explorative und konfirmatorische Analyse durch. Hat das Merkmal `Geschlecht` einen Einfluss auf die Blutdruckwerte `d_syst` und `d_diast`? Ist der t–Test anwendbar? Überlegen Sie sich weitere sinnvolle Hypothesen und Fragestellungen, und benutzen Sie dazu entsprechende statistische Verfahren.

(c) Fassen Sie Ihre Ergebnisse und Interpretationen in Form eines pdf–Dokuments mit max. 2 Seiten) zusammen.

6.) Fallbeispiel Luftschadstoffdaten (1. Teil) `grazluft.xls`; `[R 2.9, SPSS 17.0]`
Im File `grazluft.xls` finden Sie Luftschadstoff-Daten von vier Grazer Messstellen: Graz-Nord, Graz-Mitte, Graz-Ost und Graz-DonBosco in zwei Zeiträumen 16.11.2002–15.12.2002

und 1.2.2003–2.3.2003. Es sind jeweils die Tagesmittelwerte (0.00–24.00 Uhr) an Feinstaub ($PM10$), Stickstoffmonoxid (NO) und Stickstoffdioxid (NO_2) in $\mu g/m^3$ angegeben.

(a) Lesen Sie den File `grazluft.xls` von der Homepage ein. Realisierung in `R`: Speichern Sie zunächst den File `grazluft.csv` ab und lesen dann diesen File über den Befehl `read.csv2()` in `R` ein und speichern Ihn als `grazluft.dat` ab. Realisierung in `SPSS`: Definieren Sie den File `grazluft.sav`. Vergeben Sie die *Variablenlabels* wie in folgender Tabelle angegeben:

`Name`	Typ	Spalten	Dezimalen	Variablenlabel	Messniveau
`datum`	Datum	10		Datum	Metrisch
`ort`	String	14		Messort	Nominal
`pm10`	Numerisch	11	2	Feinstaub_PM10	Metrisch
`no`	Numerisch	11	2	Stickstoffmonoxid_NO	Metrisch
`no2`	Numerisch	11	2	Stickstoffdioxid_NO2	Metrisch

(b) Definieren Sie den Faktor `periode (1,2)` mit Variablenlabel *Zeitperiode* für die 2 Zeiträume 16.11.2002–15.12.2002, 1.2.2003–2.3.2003 und den Faktor `mort(1,2,3,4)` mit Variablenlabel *Messort*.

(c) Analysieren Sie die Schadstoffe mit univariaten Statistiken, Stem–and–Leaf-Plots, Histogrammen, empirischen Verteilungsfunktionen und Q-Q-Plots. Sind Auffälligkeiten in den Verteilungen zu erkennen?

(d) Vergleichen Sie die Schadstoffe bzgl. des Faktors `periode` mit Hilfe von Methoden für das Zweistichprobenproblem.

(e) Bilden Sie Box–Plot- und (Fehlerbalken)-Serien für `pm10`, `no`, `no2` getrennt, aber gemeinsam nach der Kategorie `periode` und der Gruppe `mort` (analog zu Aufgabe 4(c)).

(f) Für eine bivariate Betrachtungsweise erstelle man die Scatterplotmatrix (mit Glättungen) bezüglich `pm10`, `no`, `no2`. Gibt es bemerkenswerte Zusammenhänge mit hoher Korrelation?

Hinweise: Zusammenarbeit in Zweiergruppen ist erwünscht.

Herunter laden der Daten über die HomePage des Instituts: `www.statistics.tugraz.at`

Speichern Sie die **gesamten Übungen in einem pdf-File** mit folgendem Namen ab:
`Angstat_Nachname1*` z.B. `Angstat_Schiefer1.pdf`

und übermitteln Sie **einen File pro Gruppe** mit *Subject:* **Angstat** an die e-mail-Adresse `statistik@tugraz.at`.

Transfer der Files bis spätestens: Di. 3. 11. 2009, 16.00 Uhr

BESPRECHUNGSTERMIN: MI. 4. 11. 2009, 10.30–12.00, SR STATISTIK