1. Übungsblatt *506.556 Statistik, WS 2007/2008*

Univ.-Prof. DI Dr. Ernst Stadlober

1.) [P] Häufigkeitsdaten, diskretes Merkmal

In einer kleinen Pension wird an $n = 50$ Tagen jeweils $X =$ *Anzahl der belegten Betten* festgestellt.

```
5 7 10 8 9 9 6 7 6 9 7 8 5 4 9 7 5 9 7 6 8 7 6 7 7

6 9 6 7 9 6 5 8 6 8 7 8 9 9 9 6 8 6 5 9 7 5 7 5 7
```

(a) Zeichnen Sie ein Balkendiagramm für die absoluten Häufigkeiten.

(b) Bestimmen Sie den Modus x_{mod}, das arithmetische Mittel $\overline{x}$ und den Median x_{med}.

(c) Berechnen Sie die Streuungsmaße s, s_L, v, s_q, die Schiefmaße g_1^m, g_1^q, und die Maße für die Wölbung g_2^m, g_2^q.

2.) [C] Explorative Analyse für die Absolventenstudie aus Bsp. 1.2 `[R 2.6.0]`.

(a) Laden Sie den Datenfile `MünchnerAbsolventenstudie_1995.dat` und erstellen Sie *Säulen(Balken)- und Kreisdiagramme* der Merkmale `Note`, `Diplomarbeit` und `Engagement`.

(b) Gibt es geschlechtsspezifische Unterschiede bzgl. der Noten? Zeichnen Sie dazu die *Balken- und Kreisdiagramme* für Frauen und Männer getrennt und interpretieren Sie das Ergebnis.

(c) Unterteilen Sie die Stichprobe in zwei Schichten mithilfe einer kategorischen Variable `Zensur`. Schicht 1: Note 1 oder 2, Schicht 2: Note 3 und schlechter. Verwenden Sie dazu zum Beispiel die Funktion `cut`. (Nähere Informationen dazu erhalten Sie über den Befehl `help(cut)`.) Erstellen sie für beide Schichten das Balkendiagramm der `Studiendauer` sowie ein Balkendiagramm der `Studiendauer` gestapelt nach `Zensur`.

(d) Erzeugen Sie eine *HeatMap* sowie einen *Mosaikplot* der Faktoren `Note` und `Engagement`.

(e) Erstellen Sie die *empirische Verteilungsfunktion* der Studiendauer jeder Schicht (`Zensur`). Wie viele Semester benötigen die 25% der schnellsten Studenten in jeder Gruppe? Wie viele Semester brauchen jeweils die 25% langsamsten Studenten mindestens?
Hinweis: Das Auswählen jener Elemente der Variable `Studiendauer`, für die `Zensur==1` gilt, erhält man z.B. mit `Studiendauer[Zensur==1]`.

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

3.) [T] Eigenschaften von Lagemaßen. [Lemma 2.1.2]

(a) Für ein diskretes Merkmal X gebe man ein einfaches Beispiel an für das gilt $\overline{x} \notin \{x_1, \ldots, x_n\}$, $x_{med} \notin \{x_1, \ldots, x_n\}$, $x_i \in \mathbb{Z}$.

(b) Es sei $y_i = ax_i + b$, dann gilt:

$$\begin{aligned} y_{mod} &= ax_{mod} + b \,, \text{falls } x_i \text{ zumindest nominal} \\ y_{med} &= ax_{med} + b \,, \text{falls } x_i \text{ zumindest ordinal} \\ \overline{y} &= a\overline{x} + b \,, \text{falls } x_i \text{ zumindest intervall-skaliert} \,. \end{aligned}$$

(c) Man beweise folgende Eigenschaften.

$$\text{(i) } x_{mod} : \sum_{i=1}^{n} D(x_i, x_{mod}) = \min_z \sum_{i=1}^{n} D(x_i, z) \text{ mit } D(x_i, z) = \begin{cases} 1 & \text{für } x_i \neq z \\ 0 & \text{für } x_i = z \end{cases} ,$$

$$\text{(ii) } x_{med} : \sum_{i=1}^{n} L(x_i, x_{med}) = \min_z \sum_{i=1}^{n} L(x_i, z) = \min_z \sum_{i=1}^{n} |x_i - z| \,,$$

$$\text{(iii) } \overline{x} : \quad \sum_{i=1}^{n} Q(x_i, \overline{x}) = \min_z \sum_{i=1}^{n} Q(x_i, z) = \min_z \sum_{i=1}^{n} (x_i - z)^2 \,.$$

Hinweis zu (ii): O.B.d.A. nehme man an, dass $n = 2m + 1$ (dann ist $x_{med} = x_{(m+1)}$) und weiters dass $z > x_{(m+1)}$. Es ist zu zeigen, dass $Q(z) > Q(x_{(m+1)})$ für $z > x_{(m+1)}$.

4.) [T] Eigenschaften von Varianz und Standardabweichung. [Lemma 2.1.5]

Sei $(x_1, \ldots, x_n)$ metrisch skaliert.

(a) Man zeige, dass für alle $c \in \mathbb{R}$ gilt

$$\sum_{i=1}^{n} (x_i - c)^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 + n(\overline{x} - c)^2 \,.$$

(b) Es sei $y_i = ax_i + b$, dann gilt

$$s_y^2 = a^2 s_x^2 \quad \text{bzw.} \quad s_y = |a| s_x \,.$$

(c) Sei $E = \bigcup_{j=1}^{r} E_j$ mit $|E| = n$, $|E_j| = n_j$, $\sum_{i=1}^{r} n_j = n$, und die Stichprobe $(x_1, \ldots, x_n)$ aufgeteilt in r Schichten der Form $(x_{11}, \ldots, x_{1n_1}, x_{21}, \ldots, x_{2n_2}, \ldots, x_{r1}, \ldots, x_{rn_r})$

$$\text{Arithmetisches Mittel von Schicht } j : \overline{x}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} x_{jk} \,, \; j = 1, \ldots, r \,,$$

$$\text{Empirische Varianz von Schicht } j : s_L^2(j) = \frac{1}{n_j} \sum_{k=1}^{n_j} (x_{jk} - \overline{x}_j)^2 \,.$$

Man zeige, dass folgende Varianzzerlegung gilt

$$s_L^2 = \frac{1}{n}\sum_{j=1}^{r}\sum_{k=1}^{n_j}(x_{jk}-\overline{x})^2 = \frac{1}{n}\sum_{j=1}^{r} n_j s_L^2(j) + \frac{1}{n}\sum_{j=1}^{r} n_j(\overline{x}_j-\overline{x})^2$$

mit $\overline{x} = \frac{1}{n}\sum_{j=1}^{r} n_j\overline{x}_j$.

5.) [P] Geschichtete Stichprobe.

Für die drei Putzkolonnen einer Reinigungsfirma ergibt sich je nach Alter, Dauer der Betriebszugehörigkeit und Einsatzgebiet folgende Einkommensverteilung (in Euro) pro Monat.

1.	1645	1777	1738	1561	1769	
2.	1489	1334	1754	1311		
3.	1779	1357	1437	1517	1809	1336

Berechnen Sie für jede Putzkolonne

(a) das Durchschnittseinkommen (arithmetisches Mittel $\overline{x}_{i\cdot}$) und Median $x_{med}(i)$,

(b) die Spannweite, den interquartilen Bereich, die Standardabweichungen $s_L(i)$, $s(i)$, $s_q(i)$ und den Variationskoeffizienten $v(i)$.

(c) Wie lauten die Kenngrößen für alle 3 Putzkolonnen zusammen und die entsprechende Varianzzerlegung gemäß Aufgabe 4(c)?

6.) [C] Luftschadstoffdaten aus Bsp. 1.1 `[R 2.6.0]`
Laden Sie die Datei `Luftdaten_GrazMitte101105.dat` in `R` und führen Sie folgende Analysen durch.

(a) Analysieren Sie die Variablen `pm10` und `ltusg_k` mit den Methoden der explorativen Datenanalyse. Benutzen Sie *Histogramme, Stengel–Blatt–Diagramme (Stem–and–Leaf–Plots), Boxplots und Q–Q–Plots*. Versuchen Sie, einen einheitlichen Standard für ihre Grafiken festzulegen (Füllfarbe, Beschriftungen etc.) und geben Sie jeder Grafik einen Titel.

(b) Berechnen Sie für die Variablen `no2` und `lute` statistische Kenngrößen, die von `R` standardmäßig angeboten werden. Berechnen Sie auch s_q, g_1^q, g_2^q und v.

(c) Man erzeuge Box–Plot–Serien für die Merkmale `pm10` und `no2` getrennt bzgl. der Kategorie `monat`.

(d) Man erstelle Box–Plot–Serien für die Merkmale `pm10` und `no2` getrennt bzgl. der Kategorie `monat` und der Zeile `tag`.
Hinweis: Mit dem Befehl `par(mfrow=c(m,n))` kann man in `R` ein leeres Grafikfenster mit m Zeilen und n Spalten erzeugen, das mit den danach aufgerufenen Plots gefüllt wird.

(e) Erzeugen Sie ein *Trellis Histogramm* von `pm10` bezüglich des Faktors `tag` (drei Zeilen).

(f) Erstellen Sie die *Scatterplotmatrix* der metrischen Merkmale `pm10, lute, ltusg_k`, markiert nach dem kategorischen Merkmal `monat`. Man erzeuge einen *3D-Scatterplot* mit `lute` (x-Achse), `pm10` (y-Achse), `ltusg_k` (z-Achse).

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

Hinweise: Zusammenarbeit in Zweiergruppen ist erwünscht.
Speichern Sie Ihre Übungsaufgaben (mit entsprechenden Kommentaren) unter folgenden File–Namen ab: `Statistik_Nachname1aufgabenr.*` z.B. `Statistik_schiefer11.pdf` und übermitteln Sie die Files per e-mail mit dem Betreff **stat** an `statistik@tugraz.at`.

Transfer der Files bis spätestens: Di. 30. 10. 2007, 10.00 Uhr

Besprechungstermin: Mi. 31. 10. 2007, 16.15–17.45, HS BE01