



Prof. Dr. Antony Unwin, Dr. Ali Ünlü  
Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse  
Institut für Mathematik  
Universität Augsburg  
<http://stats.math.uni-augsburg.de/>



## Stochastik IV – Multivariate statistische Verfahren

### Übungsblatt 2

**Bearbeitung:** Dienstag 8. Mai 2007, 12.15 - 13.45 Uhr  
Raum 2001 T

1. Betrachtet wird der Datensatz **Titanic**.

- (a) Als eine Motivation für 'Mosaic Plots' zur graphischen Behandlung (hochdimensionaler) multivariater kategorialer Daten wurde in der Vorlesung genannt, dass 'Linked Highlighting' unter Benutzung von 'Barcharts' ('Spineplots') nur eingeschränkt nützlich ist. Bestimmen Sie, unter Benutzung von 'Barcharts' und 'Spineplots' mittels 'Linked Highlighting', folgende Wahrscheinlichkeiten:

$$P(\text{Survived} = \text{Yes} \mid \text{Sex} = \text{Female AND Class} = \text{First}),$$

$$P(\text{Survived} = \text{Yes} \mid \text{Sex} = \text{Female AND Class} = \text{Third}).$$

Ist ein direkter graphischer Vergleich beider Größen möglich?

- (b) Überprüfe die erhaltenen Werte mittels 'Mosaic Plots' und beschreibe einen nun möglichen direkten graphischen Vergleich der genannten Größen.
- (c) Verstehe und überprüfe Jaques Bertins Graphik über den Datensatz **Titanic** (Folie Seite 33). Nenne mögliche Vor- und Nachteile solch einer graphischen Darstellung.

2. Betrachtet wird eine Gerade

$$l : x_2 = mx_1 + b$$

im (orthogonalen) kartesischen Koordinatensystem. Angenommen sei, dass Achsen paralleler Koordinaten  $x'_1$  und  $x'_2$  derart im kartesischen Koordinatensystem platziert sind, dass die  $x'_1$ -Achse (paralleler Koordinaten) ihren Ursprung ( $x'_1 = 0$ ) in  $x_1 = 0$  (kartesischer Koordinate) hat und parallel zur  $x_2$ -Achse (kartesischer Koordinaten) verläuft.

Zeige, dass Gerade  $l$ , in parallelen Koordinaten, in eine Menge von Punkten übergeht, die sich in dem Punkt, in kartesischen Koordinaten,

$$\left( \frac{d}{1-m}, \frac{b}{1-m} \right)$$

schneiden, wenn man jeden Punkt (in parallelen Koordinaten) als (eindeutig festgelegte) Gerade im kartesischen Koordinatensystem fortgesetzt denkt. Hierbei bezeichnet  $d$  den Abstand zwischen den beiden Achsen  $x'_1$  und  $x'_2$  im kartesischen Koordinatensystem.

3. Erstellen Sie zwei 'Trellis Plots' mit der `lattice` Funktion in R, in dem die **Barley** Daten dargestellt werden. Plotten Sie die Daten mittels 'Box Plots' (Hinweis: Funktion `bwplot`). In den Spalten soll nach Jahren aufgeteilt, in den Zeilen einmal nach Sorte und einmal nach Site aufgeteilt werden.

4. Überprüfen Sie die 5 Messgrößen des **Crabs** Datensatz auf Normalität mittels:

- Histogrammen und überlagerten Dichteschätzern (R Funktion `density`),
- QQ-Plots.

Wie groß ist die jeweilige Kovarianz zwischen den 5 Variablen? Was kann man über die gemeinsame Verteilung der Messgrößen sagen?

5. Betrachtet werden die beiden Variablen *palmitoleic* und *oleic* aus dem **Olive Oils** Datensatz.

- Plotten Sie die Daten in einem Scatterplot.
- Ergänzen Sie mittels des 2-dimensionalen Dichteschätzers `kde2d` aus der MASS Library Höhenlinien für die geschätzte Dichte. Welche Bandbreiten sind für eine sinnvolle Dichteschätzung empfehlenswert?
- Ergänzen Sie nun weiter die beiden Randdichten als Projektion auf die  $x$ - bzw.  $y$ -Achse.