Prof. Dr. Katharina Morik,
Prof. Dr. Claus Weihs
Dipl.-Inform. Klaus Friedrichs,
Dipl.-Stat. Julia Schiffner,
Dr. Issam Ben Khediri

Dortmund, 24.04.12

Abgabe: bis Mi, 02.05., 12.00 Uhr an
schiffner@statistik.tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2012

Blatt 3

Aufgabe 3.1 (2 Punkte)

Im Jahr 2011 wurden in Europa 606.000 gefälschte Euro-Scheine eingezogen. Insgesamt befinden sich ca. 14 Milliarden Scheine im Umlauf. 20% der von der EZB in Umlauf gebrachten (echten) Scheine sind 20-Euro-Scheine. Unter Geldfälschern ist der 20-Euro-Schein sehr beliebt. 36% der sichergestellten Fälschungen sind 20-Euro-Noten.

1. Mit welcher Wahrscheinlichkeit ist eine beliebige Euro-Banknote ein 20-Euro-Schein?
2. Sie erhalten einen 20-Euro-Schein. Mit welcher Wahrscheinlichkeit handelt es sich um eine Fälschung?

Aufgabe 3.2 (5 Punkte)

Auf der Homepage liegt der bekannte Schweizer-Banknoten-Datensatz `bank.txt`. Er enthält die Ergebnisse von Längenmessungen an 200 Schweizer 1000-Franc-Scheinen (100 echten und 100 gefälschten) in der Einheit mm. Welche Längen genau gemessen wurden, können Sie der Datei `bank_info.txt` sowie der Abbildung unten entnehmen. Lesen Sie die Daten mit der Software Ihrer Wahl ein (in R: `read.table`).

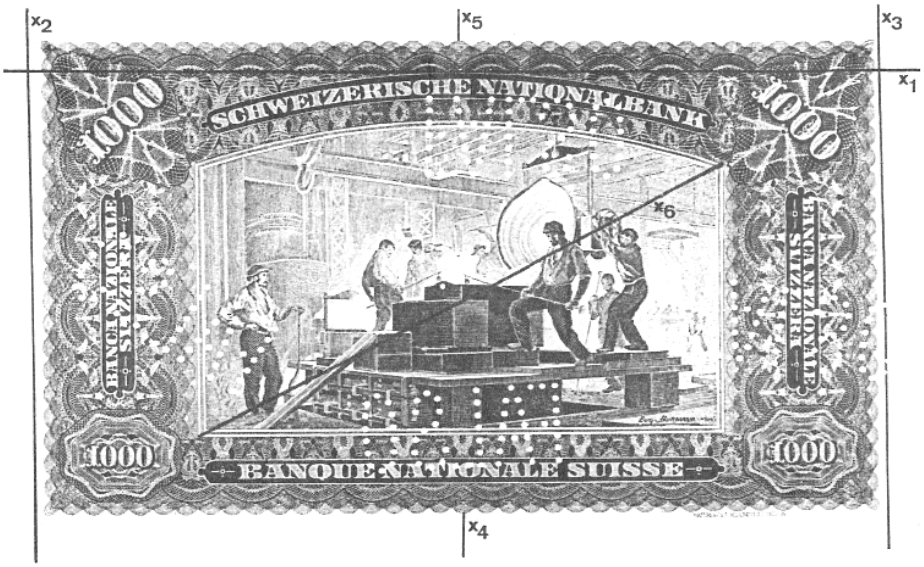

Figure 1.1 An old Swiss 1000-franc bill

Quelle: B. Flury & H. Riedwyl, Multivariate Statistics: a Practical Approach, Chapman and Hall, London, 1988

1. Zeichnen Sie die Scatterplot-Matrix für die 6 gemessenen Längen (R-Funktion `pairs`). Färben Sie dabei die Punkte nach der Klassenzugehörigkeit ein.

2. Berechnen Sie die folgenden Kennzahlen für die 6 gemessenen Längen: arithmetisches Mittel, Median sowie die Standardabweichung. In R sind die Funktionen `summary` und `sd` hilfreich.

3. Betrachten Sie die Variable `Inner.Frame.upper`. Kann man annehmen, dass diese Variable für die beiden Gruppen, echte und gefälschte Geldscheine, jeweils normalverteilt ist? Sie können z.B. folgendermaßen vorgehen: Splitten Sie `Inner.Frame.upper` nach echten und gefälschten Geldscheinen auf. Standardisieren Sie die Variable jeweils auf Mittelwert 0 und Varianz 1 (R-Funktion `scale`) und berechnen Sie anschließend die 0.05, 0.1, 0.15, ... , 0.9, 0.95-Quantile (R-Funktion `quantile`). Vergleichen Sie jeweils die berechneten Quantile mit den theoretischen Quantilen der eindimensionalen Normalverteilung mit Erwartungswert 0 und Varianz 1 (R-Funktion `qnorm`).

4. Ermitteln Sie mithilfe eines 2-Stichproben t-Tests (R-Funktion `t.test`), ob sich echte und gefälschte Geldscheine in der Variable `Inner.Frame.upper` signifikant unterscheiden (für $\alpha = 0.05$).

Aufgabe 3.3 (4 Punkte)
Die Deutsche Bundesbank gibt jährlich Statistiken zum Falschgeldaufkommen in Deutschland heraus. In der Datei `euro.txt` auf der Homepage finden Sie für die Jahre 2003 bis 2011 die Anzahlen der sichergestellten gefälschten Banknoten sowie den verursachten finanziellen Schaden in Mio. Euro.

1. Lesen Sie die Daten ein und passen Sie ein lineares Regressionsmodell der Form $y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$, $i = 1, \ldots, 9$, an die Daten an, das die Schadenssumme aus der Anzahl der gefälschten Banknoten vorhersagt (R-Funktion `lm`).

2. Tragen Sie in einem Diagramm die Schadenssumme gegen die Anzahl der gefälschten Banknoten auf (R-Funktion `plot`) und zeichnen Sie die Regressionsgerade ein (R-Funktion `abline`).