Bernd Klaus (bernd.klaus@imise.uni-leipzig.de)
Verena Zuber (verena.zuber@imise.uni-leipzig.de)

http://uni-leipzig.de/~zuber/teaching/ws10/r-kurs/

Dieses Übungsblatt ist dem Datensatz aus der Veröffentlichung *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression* von Golub et al. gewidmet.

- Genexpressionsdaten mit $n = 38$ Beobachtungen und $p = 3051$ Genen
- Faktorvariable `golub.cl` beschreibt welche Form von Leukämie ($k = 2$) bei der entsprechenden Beobachtungen vorliegt
  - ALL: acute lymphoblastic leukemia ($n_0 = 27$)
  - AML: acute myeloid leukemia ($n_1 = 11$)
- Die Matrix `golub.gnames` enthält Information zu den Bezeichnungen der beobachteten Gene
- Die Daten sind schon praeprozessiert und zu finden als `RData`-File im Netzwerkordner unter L:\R-Kurs oder im `R`-Paket `multtest` unter *Data*

1 Aufgabe: Deskriptives

(a) Lesen Sie den Datensatz mit Hilfe des `load()`-Befehls ein und überprüfen Sie welche Information das `RData`-File enthält und die Dimension der Genexpressions-Daten `golub`.

(b) Berechnen Sie für jedes Gen die Spannweite (Maximum-Minimum). Betrachten Sie die Summary und erstellen Sie ein Histogramm der Spannweiten.

(c) Berechnen Sie für jedes Gen die Varianz. Greifen Sie die 10 Gene mit der höchsten Varianz heraus, erstellen Sie für diese je einen Boxplot und fassen Sie die Einzelboxplots in einer Graphik zusammen.

2 Aufgabe: Differentielle Expression

(a) Berechnen Sie für jedes Gen den Fold Change (`ALL-AML`).

(b) Wie viele Gene sind differentiell unterexprimiert? Erstellen Sie eine Liste mit den differentiell unterexprimierten Genen (negativer Fold Change).

(c) Berechnen Sie für jedes Gen den t-Score.

(d) Erstellen Sie zwei Ranglisten einmal nach dem Fold-Change einmal nach dem t-Score. Vergleichen Sie die Top20 der beiden Listen. Welche Gene tauchen in beiden auf?

(HINWEIS: Nutzen Sie zur Lösung die `match` Funktion.)

(e) Erstellen Sie ein Histogramm der t-Scores. Beschreiben Sie die Form.

(f) Berechnen Sie den Mittelwert und die Standardabweichung für die 90% der mittleren absoluten t-Scores. Fügen Sie in das Histogramm eine Normalverteilung mit diesen Parametern ein. Was fällt Ihnen insbesondere bei den niedrigsten und höchsten t-Scores auf?