

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

<http://uni-leipzig.de/~zuber/teaching/ws12/r-kurs/>

## 1 Aufgabe: Testen an PISA-Daten

Das *Programme for International Student Assessment*, kurz PISA, ist eine standardisierte Bewertung von (15 jährigen) Schülern unter den teilnehmenden Staaten. Ziel der Regierungen ist, eine Datenbasis zur länderübergreifenden Forschung zu ermöglichen. Im Datensatz *PISA.csv* finden Sie die Ergebnisse einiger ausgewählter OECD-Staaten, getrennt nach dem Geschlecht (Variable *sex*: 1 Female, 2 Male, *Perc\_Sex* gibt den Anteil an). Folgende Variablen sind von Interesse:

- *R00 - R06*: Mittlerer Score zur Lesekompetenz im Jahr 2000 bzw. 2006
- *M00 - M06*: Mittlerer Score zur Kompetenz in der Mathematik im Jahr 2000 bzw. 2006
- *S00 - S06*: Mittlerer Score in den Naturwissenschaften (*science*) im Jahr 2000 bzw. 2006

1. Laden Sie den Datensatz *PISA.csv* von der Homepage herunter und lesen sie ihn ein.
2. Untersuchen Sie deskriptiv, ob sich die drei PISA-Scores des Jahres 2006 im Vergleich zum Jahr 2000 verändert haben. (Gehen Sie hierbei und im weiteren nicht näher auf irgendwelche Geschlechtsunterschiede ein)
3. Untersuchen Sie mit einem geeigneten Test, ob sich die drei PISA-Scores signifikant verändert haben.

## 2 Aufgabe: Google-Suche in Deutschland

Im Datensatz "google.csv" finden Sie getrennt nach dem Bundesland die absolute Suchhäufigkeit nach:

- "Krise" im Jahr 2008 (Krise08)
- "Krise" im Jahr 2009 (Krise09)
- "Arbeitsamt" im Jahr 2010 (Arbeitsamt)

Dazu enthält der Datensatz einen Ost-West Indikator (OW) (Berlin wird von dieser Kategorisierung ausgenommen).

- a) Lesen Sie den Datensatz ein und wandeln Sie die Variable "OW" in einen Faktor um, der fehlende Wert zulässt und mit 0 West und 1 Ost kodiert.
- b) Führen Sie den geeigneten *t*-Test durch, um zu untersuchen, ob nach dem Begriff "Krise" im Jahr 2008 stärker gesucht wurde als im Jahr 2009. Erstellen Sie zusätzlich einen Boxplot und führen Sie den entsprechenden nonparametrischen Test durch. Interpretieren Sie die Ergebnisse.

- c) Überprüfen Sie mittels des Ost-West Indikator, ob diese Abnahme der Suche nach dem Begriff "Krise" im Osten stärker ausgeprägt ist als im Westen.
- d) Untersuchen Sie die Hypothese, dass im Osten häufiger nach dem Begriff "Arbeitsamt" gesucht wird als im Westen. Begründen Sie Ihre Entscheidung.

### 3 Aufgabe: Zu den Annahmen des $t$ -Test

1. Ziehen Sie 100 exponential-verteilte Zufallszahlen mit dem Parameter  $\lambda = 0.1$  und speichern Sie diese in dem Objekt `X1`. Erstellen Sie analog ein Objekt `X2` von 100 Zufallszahlen, für die Sie erneut 100 exponential-verteilte Zufallszahlen `HX2` mit dem Parameter  $\lambda = 0.1$  ziehen und anschließend alle Elemente von 20 subtrahieren. Ein Element  $i$  des Vektors `X2` berechnet sich also als `20-HX2[i]`.
2. Zeichnen Sie für die beiden Objekte in einer Graphik je den Kerndichteschätzer.
3. Führen Sie den  $t$ -Test durch, um zu untersuchen, ob die beiden Objekte unterschiedliche Mittelwerte besitzen.
4. Führen Sie den Wilcoxon Rangsummen Test durch, um zu untersuchen, ob die beiden Objekte unterschiedliche Mediane besitzen.

### 4 Aufgabe: Produktionskontrolle bei Hustensaft

Laut Produktbeschreibung enthält der Hustensaft "Mikasolvan" 40g/Liter des Wirkstoffes "Halsruhe". Die Produktion des Saftes wird angehalten, wenn die Konzentration des Wirkstoffes deutlich zu niedrig ist.

Um die Produktion zu überprüfen, wurden 9 Flaschen zufällig entnommen und die Konzentration von "Halsruhe" gemessen. Die Messwerte finden Sie in der Datei `Hustensaft.csv` auf der Homepage.

Begründen diese Daten einen Produktionsstopp?