Statistik I — bis jetzt

K 1 Einführung

K 2 Beschreibende Statistik

K 3 Graphiken

K 4 Schätzen

K 5 Testen

und jetzt zurück zu Graphiken

K 3.5 Interaktive Graphiken

K 3.5 Interaktive Graphiken

3.5.1 Beispiel: Die Patienten von Dr. H. Shipman

- Daten aus dem offiziellen Untersuchungsbericht
- Tode von 508 Patienten zwischen 1973 und 1998
- Alter, Name, Todestag, Todesort, Entscheidung der juristischen Untersuchung
- (keine Information zur Todeszeit oder
  ob der Arzt anwesend war)

3.5.2 Interaktive Graphik und EDA

EDA (Explorative Datenanalyse) bedeutet die Generierung von Hypothesen — nicht die Überprüfung.

IG ist (idealerweise) die direkte Manipulation von statistischen Objekten in Graphiken. Statistische Objekte sind, z.B. Datenpunkte, Säulen in Säulendiagrammen oder Histogrammen, Achsen. Sie werden aus graphischen Objekten (z.B. Pixeln oder Linien) erstellt.

IG muss schnell, flexibel und “forgiving” (verzeihend) sein.

Für IG können Sie die Software MONDRIAN
http://stats.math.uni-augsburg.de/Mondrian/
von Martin Theus verwenden.

3.5.3 Hauptmerkmale von Interaktive Graphik

- Abfrage
  Welche Werte werden dargestellt?
  Default und erweiterte Abfragen.

- Selektion und Linking (Verknüpfungen)
  Ausgewählte Fälle werden überall hervorgehoben.
  (1) Bei Histogrammen und Säulendiagrammen sieht man #(selektierte aus X), bei Spineplots sieht man P(selektiert|X).
  (2) Gegeben $X = x_j$ selektiert, wird $X = x_j | Y = y_k$ in der $Y = y_k$ Säule gezeigt.

- Umformatierung
  Fenstergröße ändern, Seitenverhältnis ändern, Punkte vergrößern, neuskalieren (Minimum, Maximum, Binbreite), gemeinsame Skalierung, sortieren, zoomen, ...

- Multiple Ansichten
  Es gibt keine optimale Darstellung. Verschiedene Darstellungen liefern verschiedene Einsichten.

3.5.4 Interaktivität und statistische Graphiken

- Säulendiagramme (und Spineplots)
- Histogramme (und Spinogramme)
- Boxplots
- Streudiagramme

K 3.6 Multivariate Graphiken

3.6.1 Parallelkoordinatenplots (für stetige Variablen)

Um viele Variablen gleichzeitig darzustellen. Matrizen von Streudiagrammen sind nur für ein Paar Variablen brauchbar. Parallele Koordinaten Plots sind von Al Inselberg in den achziger Jahren vorgeschlagen worden.

- Jede Variable hat ihre eigene vertikale Achse. Als Default wird der Wertebereich auf $[0, 1]$ transformiert.

- Jeder Fall wird durch einen Polygonzug mit Knoten auf den Achsen bei den entsprechenden Werten dargestellt.

- Die Skalierung und die Reihenfolge der Achsen üben einen wesentlichen Einfluss auf den Plot aus.

Beispiel von Parallelkoordinatenplots:
Zehnkampf (individuelle jährliche Bestleistungen)

- Für alle 10 Disziplinen gibt es die tatsächliche Leistung (in Sekunden oder Metern) und die dafür vergebene Punktzahl. (Quelle: www.decathlon2000.ee)

- Es gibt 7968 Fälle, von 1985 bis 2006, alle haben mindestens 6800 Punkte erreicht.

- Mögliche graphische Darstellungen:
  - Histogramme
  - Boxplots
  - Streudiagramme
  - Parallelkoordinatenplots

Ziele bei der Analyse der Zehnkampfdaten

- Waren die Gewinner in jedem Disziplin die Besten?
- Gab es außerordentliche Leistungen in einigen Disziplinen?
- Wie sind die Leistungen in den Disziplinen verteilt?
- Sind die Leistungen in einigen Disziplinen hoch korreliert?
- Welche Disziplinen waren ausschlaggebend?
- Sind die Punkte gleich wert?
- Haben sich die Leistungen über die Jahre gesteigert?

Interaktive Optionen für Parallelkoordinatenplots

- Abfrage
- Selektion — Punkte/Linien
- Skalierung
  - Invertierung
  - gemeinsame Skalierung
  - Ausrichtung nach Statistiken oder nach Fällen
- Neuordnen und Sortierung von Variablen
  - per Hand
  - nach der Reihenfolge der Variablen
  - nach Statistiken der Variablen (für ausgewählte Daten)
- Darstellungsmöglichkeiten
  - alphablending
  - Hotselektion
  - Stutzen
  - Boxplots

Beispiel von Mosaicplots:
Der Rochdale Datensatz

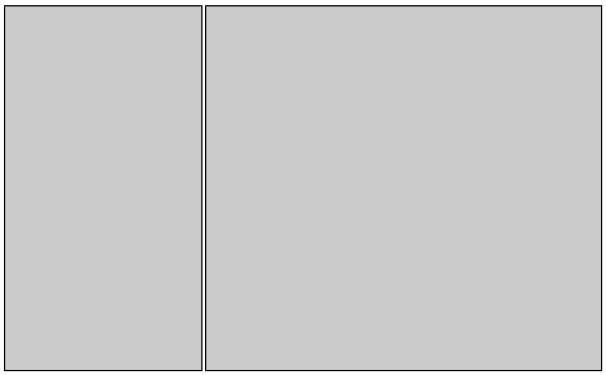
665 Haushalte aus Rochdale, England, haben an einer Umfrage teilgenommen. Unter anderem wird die Information in 8 binären Variablen zusammengefasst:

Variablename	**Bedeutung**
wEcon	Frau arbeitet/nicht
husUnemp	Mann arbeitslos/nicht
c4	Kind ≤ 4
Wife38	Frau älter als 38/nicht
WifeEdx	Frau hat eine Ausbildung/nicht
HusEdx	Mann hat eine Ausbildung/nicht
Asi	Familie aus Asien
OHwork	Andere im Haushalt arbeiten/nicht

3.6.2 Mosaic Plots (für kategorielle Variablen)

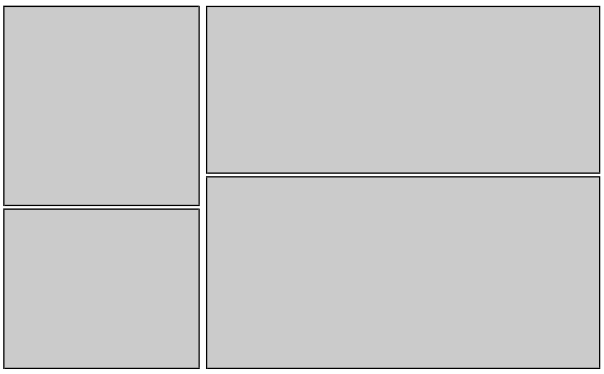
Mosaicplots stellen multivariate kategorielle Kombinationen in Rechtecken dar. Deren Größe zeigt (normalerweise) die Anzahl und deren Lage spiegelt die Kombination wider.

1. Zuerst wird die horizontale Achse nach der ersten Variable aufgeteilt (so daß ein eindimensionales Mosaic Plot einem Spine Plot gleicht).

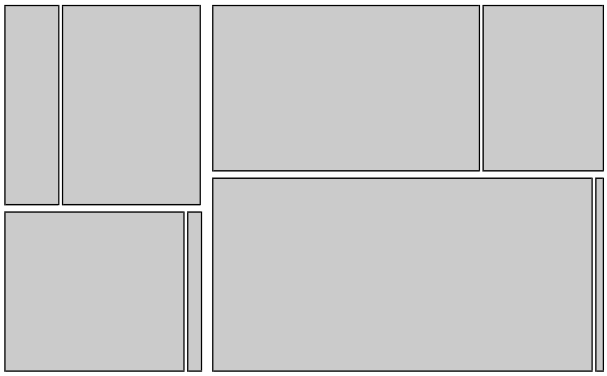
   Frau arbeitet/nicht.

2. Dann wird jede Spalte vertikal nach der zweiten Variable aufgeteilt. Daraus ergeben sich den zweidimensionalen Kombinationshäufigkeiten entsprechende Rechtecke.



Frau arbeitet/nicht, älter als 38/nicht.

3. Diese Rechtecke können im Prinzip immer weiter horizontal und vertikal aufgeteilt werden. Ein Mosaic Plot für acht binäre Variablen hat höchstens $256 (= 2^8)$ Rechtecke. Der Übersicht wegen werden leere Zellen (im Datensatz nicht vorhandenen Kombinationen) mit einem roten Strich markiert.



Frau arbeitet/nicht, älter als 38/nicht, Kind unter 4/nein

Mosaic Plot Variationen

- Anzahl
- Erwartete
- Gleiche Bingröße
- Multiple Säulendiagramme
- rmb Plots (’Relative Multiple Barcharts’)
- Fluctuationsdiagramme
- Gewichtete Mosaicplots
- Doubledeckerplots

Interaktive Optionen für Mosaicplots

- Abfrage
- Selektion
- Anzahl der Variablen
- Reihenfolge der Variablen
- Reihenfolge der Kategorien innerhalb der Variablen
- Variationen (einschließlich Größe und Seitenverhältnis)