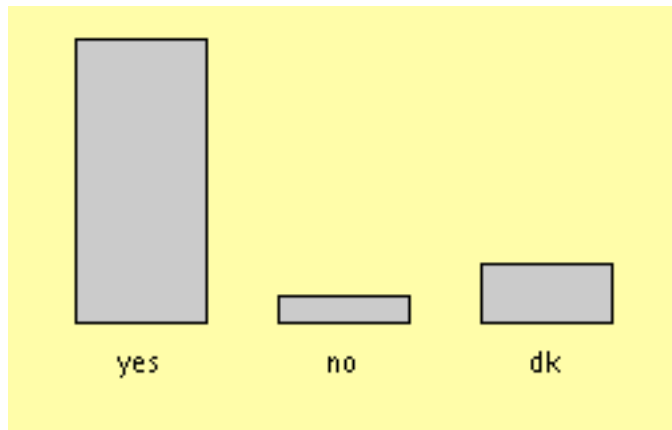
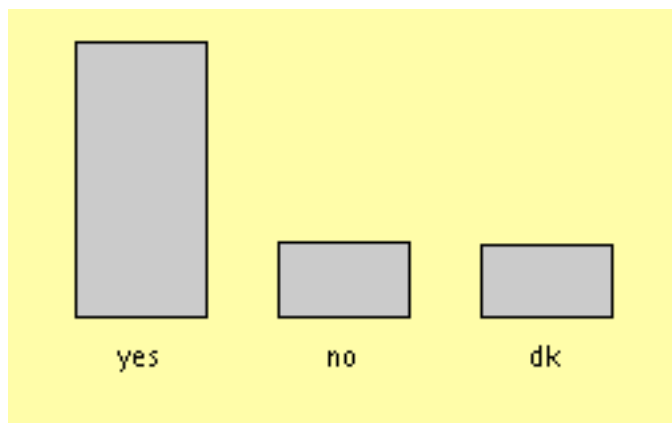


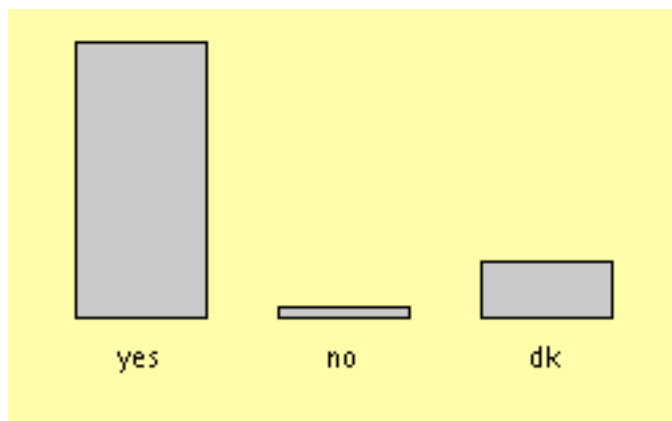
## Slovenische Umfrage Grafiken



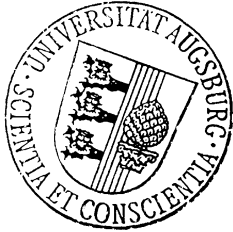
Unabhängigkeit



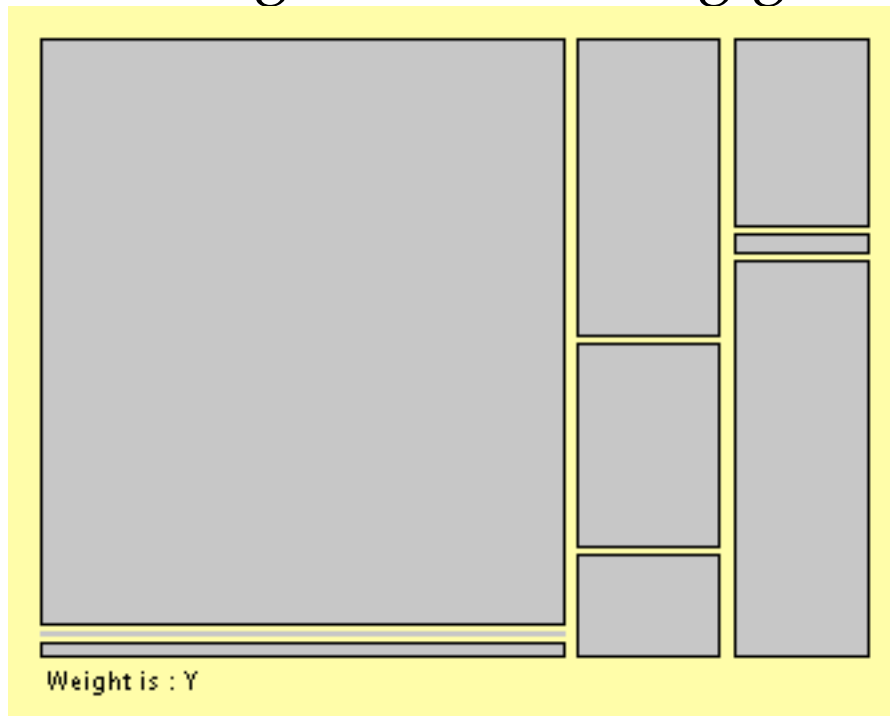
Trennung



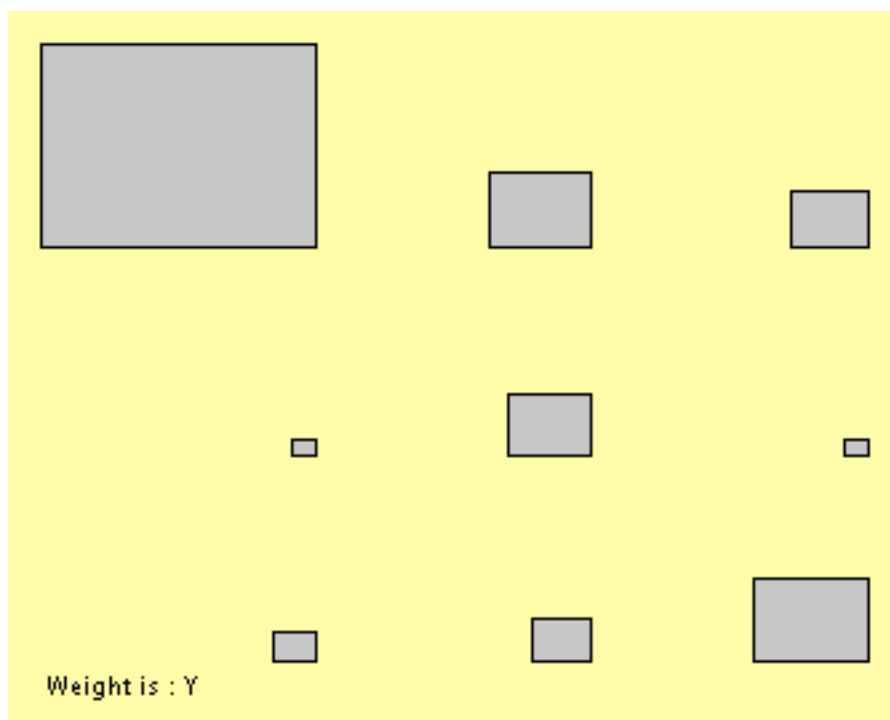
Abstimmen



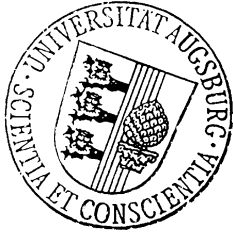
# Trennung und Unabhängigkeit



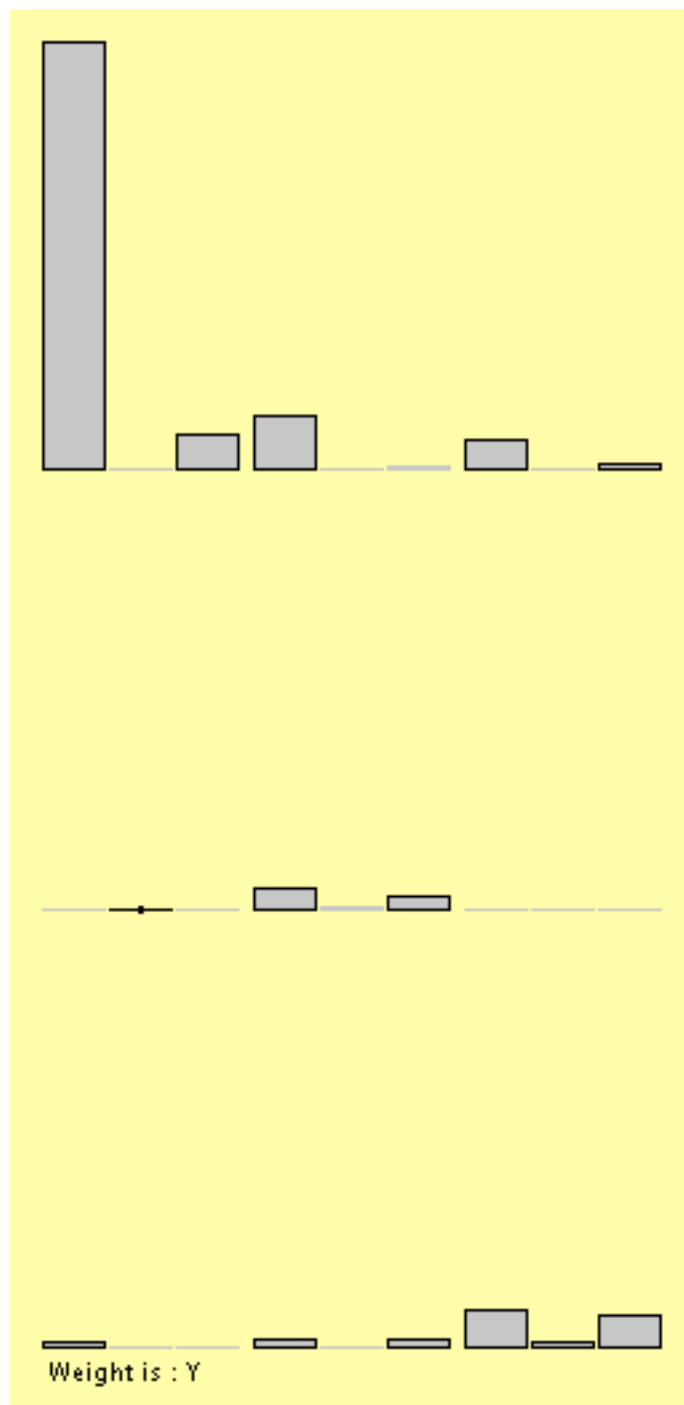
Mosaic Plot



Fluctuation



# Trennung, Unabhängigkeit, Abstimmen (Multiple Barcharts)



## Zerlegung von $X^2$ und $G^2$

---

$$(a) X^2 = \sum_i \frac{(n_i - m_i)^2}{m_i}$$

Jede Zelle  $i$  macht einen positiven Beitrag zur Statistik

$$= (\text{das standardisierte Residuum})^2$$

so dass  $X^2$  für Residuenanalysen sich eignet.

$$(b) G^2 = 2 \sum_i n_i \log \frac{n_i}{m_i}$$

oder eher  $G^2 = -2(l_m - l_s)$  für Modell M.

wo  $l_s$  das Loglikelihood für das saturierte Modell ist.

## $G^2$ ist besser als $X^2$ , um Modelle zu vergleichen

Nehmen wir an, wir haben zwei Modelle  $M_2$  und  $M_1$  mit Devianzen und Freiheitsgraden  $G^2(M_2)$ ,  $\nu_2$  bzw.  $G^2(M_1)$ ,  $\nu_1$ , wobei  $M_2$  ein Spezialfall von  $M_1$  ist mit  $\nu_2 > \nu_1$  und

$$G^2(M_1) \leq G^2(M_2)$$

Um  $M_1$  und  $M_2$  zu vergleichen, benutzen wir

$$\begin{aligned} -2(l_{m_2} - l_{m_1}) &\sim \chi_{\nu_2 - \nu_1}^2 \\ &= G^2(M_2) - G^2(M_1) \\ &= G^2(M_2|M_1) \end{aligned}$$

Asymptotisch gilt

$$(i) G^2(M) \sim \chi_\nu^2$$

und

$$(ii) G^2(M_2|M_1) \sim \chi_{\nu_2-\nu_1}^2$$

Es kann gezeigt werden, dass die Asymptotik für (ii) viel besser ist. Die Zerlegung von  $G^2$  für eine Hierarchie von Modellen hilft uns bei der Modellauswahl:

z.B.  $n = 3$

$$\begin{aligned} G^2(X, Y, Z) &= G^2(X, Y, Z) - G^2(X, YZ) \\ &\quad + G^2(X, YZ) - G^2(XZ, YZ) \\ &\quad + G^2(XZ, YZ) - G^2(XY, XZ, YZ) \\ &\quad + G^2(XY, XZ, YZ) - G^2(XYZ) \\ &= G^2[(X, YZ)|(X, Y, Z)] \\ &\quad + G^2[(XZ, YZ)|(X, YZ)] \\ &\quad + G^2[(XY, XZ, YZ)|(XZ, YZ)] \\ &\quad + G^2(XY, XZ, YZ) \end{aligned}$$

## **Modellauswahl**

---

### **1. Vorwärts**

Man fängt mit dem Unabhängigkeitsmodell an und fügt schrittweise kompliziertere Interaktionen hinzu, bis  $G^2$  nicht mehr signifikant ist.

### **2. Rückwärts**

Man fängt mit dem saturierten Modell an und lässt in jedem Schritt eine Interaktion wegfallen bis  $G^2$  signifikant ist.

(In beiden Fällen wird/kann man mehrere Modelle finden, die akzeptabel sind.)

## Leere Zellen und Zellen mit kleinen Häufigkeiten

---

$n$  = Anzahl der Fälle

$N$  = Anzahl der Zellen

$\frac{n}{N}$  ist klein wenn (a)  $n$  klein oder (b)  $N$  groß ist.

Für das Nullmodell mit  $\pi = \frac{n}{N}$  wurden wir  $N e^{-\frac{n}{N}}$  leere Zellen erwarten.

z.B. Für  $n = 1000$ ,  $N = 200$  gilt  $N e^{-\frac{n}{N}} \approx 1.35$

Wir haben aber nie das Nullmodell.

z.B. in der Slovenischen Umfrage  $n = 2076$  und  $N = 27$  und trotzdem finden wir eine leere Zelle.

Oder im Rochdale Datensatz, wo  $n = 665$  und  $N = 256$ , haben wir 165 leere Zellen.

## **Strukturelle Nullen oder Stichproben Nullen?**

Wenn leere Zellen strukturelle Nullen sind – z.B. keine überlebende Kinder in der Mannschaft im Titanic Datensatz – müssen wir diese Zellen ignorieren und dementsprechend Freiheitsgrade wegnehmen.

Wenn leere Zellen nicht strukturelle Nullen sind, dann kann es zu Anpassungsproblemen kommen. So lange die suffizienten Statistiken  $> 0$  sind (die  $n$  in  $X_n = X_m$ ) kann man das Modell schätzen. Das saturierte Modell ist natürlich dann nicht schätzbar.

## Logistische Regression und loglineare Modellen

Für loglineare Modelle müssen alle Variablen kategoriell sein. In der logistischen Regression muss die abhängige Variable binär sein.

z.B.  $N = 2 \times J \times K$  Zellen und  $n$  Fälle

### **Logistische Regression**

Es gibt  $JK$  Paare zu modellieren und die Summen der  $JK$  Paare sind festgelegt. Das saturierte Modell hat  $JK$  Parameter.

### **Loglinear Modelle**

Es gibt  $2JK$  Zellen zu modellieren. Das saturierte Modell hat  $2JK$  Parameter.

### **Logistische Regression**

$$P(Y = 1 | j, k) = \frac{1}{1 + e^{-\dots}}$$

Das Modell ohne Interaktion hat LogOdds der Form:

$$\log \frac{m_{1jk}}{m_{2jk}} = \alpha + \beta_j^{X_2} + \beta_k^{X_3}$$

## Loglinear Modelle

$$\begin{aligned} E[Y = 1 \text{ und } j \text{ und } k] &= m_{1jk} \\ &= e^{\dots} \end{aligned}$$

Das Modell  $(X_1X_2, X_1X_3, X_2X_3)$  hat LogOdds der Form:

$$\begin{aligned} \log \frac{m_{1jk}}{m_{2jk}} &= [\mu + \lambda_1^{X_1} + \lambda_j^{X_2} + \lambda_k^{X_3} \\ &\quad + \lambda_{1j}^{X_1X_2} + \lambda_{1k}^{X_1X_3} + \lambda_{jk}^{X_2X_3}] \\ &\quad - [\mu + \lambda_2^{X_1} + \lambda_j^{X_2} + \lambda_k^{X_3} + \dots] \\ &= (\lambda_1^{X_1} - \lambda_2^{X_1}) + (\lambda_{1j}^{X_1X_2} - \lambda_{2j}^{X_1X_2}) \\ &\quad + (\lambda_{1k}^{X_1X_3} - \lambda_{2k}^{X_1X_3}) \end{aligned}$$

aber  $\lambda_2^{X_1} = -\lambda_1^{X_1}$      $\lambda_{2j}^{X_1X_2} = -\lambda_{1j}^{X_1X_2}$  ...

so dass  $\log \frac{m_{1jk}}{m_{2jk}} = 2\lambda_1^{X_1} + 2\lambda_{1j}^{X_1X_2} + 2\lambda_{1k}^{X_1X_3}$

(N.B. Der Interaktionsterm  $\lambda_{jk}^{X_2X_3}$  erscheint nicht im logistischen Modell. Die LogOdds aus dem Modell  $(X_1X_2, X_1X_3)$  hätten auch diese Form, aber wir brauchen die  $X_2X_3$  Terme, um die Zellenpaar-Summen festzuhalten. Sonst wäre der Vergleich mit der logistischen Regression nicht möglich.)

## Graphische Modelle

---

Gegeben seien  $m$  Variablen, sowie ein hierarchisches log-lineares Modell  $M$ . Der Graph  $(V,E)$  mit Knotenmenge  $V$  und Kantenmenge  $E \subseteq V \times V$  heißt zu  $M$  gehöriges Graphisches Modell, wenn

- $V$  genau  $m$  Knoten enthält, die mit den  $m$  Variablen von  $M$  identifiziert werden;
- für jede in  $M$  enthaltene Zweifach-Interaktion gibt es eine entsprechende Kante in  $E$ ;
- für jeden vollständigen Subgraphen der Mächtigkeit  $r$  die entsprechende  $r$ -fach Interaktion in  $M$  enthalten ist.

Falls die dritte Bedingung nicht erfüllt werden kann, heißt das zugrundeliegende Modell nichtgraphisch.

# **Eigenschaften Graphischer Modelle**

---

## **1. Bedingte Unabhängigkeiten**

Seien  $A$  und  $B$  Teilmengen der Variablenmenge  $V$ . Dann gilt:  $A$  und  $B$  sind bedingt unabhängig gegeben  $C$  genau dann, wenn alle Pfade von Elementen in  $A$  zu Elementen von  $B$  Elemente aus  $C$  enthalten.

## **2. Zerlegbarkeit der ML-Gleichungen**

Ein graphisches Modell besitzt zerlegbare ML-Gleichungen, wenn der zugehörige Graph keine sehnenslosen Kreise der Länge größer oder gleich vier besitzt. Solche Graphen sind triangulierbar.

## **3. Kollabierbarkeit von Modellen**

Ein graphisches Modell ist kollabierbar über eine Menge  $A$  von Variablen, falls die Nachbarschaften der Zusammenhangskomponenten von  $A$  jeweils vollständige Teilgraphen sind.

# Graphische Darstellungen für Datensätze mit mehreren kategoriellen Variablen

---

## 1. verknüpfte Säulendiagramme

Ohne Verknüpfungen werden nur die univariaten marginalen Verteilungen gezeigt. Mit Verknüpfungen kann man einige Strukturen erkennen, aber die gehighlighteten Proportionen sind schwer zu beurteilen. Man braucht komplizierte Auswahlmöglichkeiten, um multivariate Strukturen zu untersuchen (u.a. Schnittmenge).

### **Vorteil:**

Die Säulenhöhen spiegeln die absoluten Zahlen wider.

### **Nachteil:**

Gehighlightete Selektionen sind nicht vergleichbar.

## 2. Spineplots

Die Säulenfläche stellt die Anzahl dar (wie in einem Säulendiagramm), aber jetzt wird die Höhe (statt die Breite) in jeder Kategorie gleich.

### **Vorteil:**

Proportionen sind direkt nach Höhen vergleichbar.

**Nachteil:** Absolute Zahlen sind nicht vergleichbar.

### 3. **Mosaic Plots**

Eine multivariate Verallgemeinerung von Spineplots.

Zuerst wird die X-Achse nach der ersten Variable in Spalten zerlegt. Dann wird jeder dieser Säulen vertikal nach der zweiten Variable zerlegt. Mit der dritten Variable werden die so erzeugten Rechtecke nochmals einzeln horizontal zerlegt u.s.w.

Die Größe jeder der aus diesem Prozeß erstellte Fläche soll der Anzahl für diese Kombination entsprechen.

#### **Vorteile:**

Alle Kombinationen können dargestellt werden.

Neue Variablen Reihenfolgen bringen neue Einsichten.

Interaktives Abfragen bringt Verständnis.

Verknüpfungen mit anderen Darstellungen sind möglich.

#### **Nachteile:**

Schwer, leere Zellen/Kombinationen darzustellen.

Andere Variablen Reihenfolgen geben andere Eindrücke.

Struktur kann schwer zu entziffern sein.

## **Warum soviel über loglineare Modelle?**

---

- Anwendungen (aber nur unbefriedigende analytische Methoden)
- Viele Daten aber wenig Information
- Definitionen der Kategorien haben viel Einfluß (z.B. alt/jung). Es kann schwer sein, Fälle zu klassifizieren.
- Neue Forschung: graphische Modelle
- Neue Forschung: Mosaic Plots