Statistik für Digital Humanities

Kategorische Variablen

Dr. Jochen Tiepmar

Institut für Informatik
Computational Humanities
Universität Leipzig

28. Juni 2021

[Letzte Aktualisierung: 28/06/2021, 08:49]

Wiederholung Datenskalierung

- Kategorische Skalierung
  - $\rightarrow$ Binär & Nominal
  - $\rightarrow$ Ordinal
- Numerische Skalierung
  - $\rightarrow$ Intervall
  - $\rightarrow$ Absolut

Wiederholung Binär & Nominal

- Eigenschaften wie "krank" – "gesund", "Raucher", "Nichtraucher", Geschlecht, Farben, Berufsgruppe, Tierart, Apfelsorte
- jede Beobachtung einer Merkmalsausprägung wird genau einer bestimmten Klasse (Kategorie) zugeordnet
- Klassen können nicht geordnet sondern nur unterschieden werden
- Klassen auch z.B. durch natürliche Zahlen oder Buchstaben charakterisiert
- Binär: 2 Kategorien (Biologisches Geburtsgeschlecht)

Statistik mit Kategorischen Variablen

Bisher:

- Intervall- und Absolutskaliert $\rightarrow$ Parametrische Verfahren
- Ordinalskaliert $\rightarrow$ Ranking (Vorlesung Nichtparametrische Testverfahren)
- Nominal $\rightarrow$ Diese Vorlesung

- 2 Variablen $\rightarrow$ Pearsons χ^2 Chi Quadrat, Fishers Test
- Mehr als 2 Variablen $\rightarrow$ Log-Lineare Analyse

Statistik mit Kategorischen Variablen

- Statt Mittelwerten verwenden wir jetzt Häufigkeiten
- **Kontingenztabelle** Contingency Table, Cross Tabulation, Crosstab

Beispiel : Können wir Katzen tanzen beibringen?

		Belohnung Leckerli	Belohnung Lob	Insg
Tanzen sie?	Ja	28	48	76
	Nein	10	114	124
	Insg	38	162	200

Pearsons χ^2 Test

Pearson, K. (1900): *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling*
Fisher, R.A. (1922): *On the interpretation of chi square from contingency tables, and the calculation of P*

- Grundidee: Berechne Abweichung zwischen beobachteten Werten und den zufällig zu erwartenden
- $Abweichung = \sum (Beobachtet - Modell)^2$
- Normalisierung ergibt: $\chi^2 = \sum \frac{(Beobachtet_{ij} - Modell_{ij})^2}{Modell_{ij}}$
- $Modell_{ij} = \frac{Zeilensumme_i * Spaltensumme_j}{n}$ zu erwartende Werte
- $H_0 =$ Es gibt keine signifikante Beziehung zwischen den Variablen
- $df = (Spalten - 1) * (Zeilen - 1)$
- $\chi^2 > \chi^2_{kr} \rightarrow H_0$ kann verworfen werden

Pearsons χ^2 Test

$\chi^2 = \sum \frac{(Beobachtet_{ij} - Model_{ij})^2}{Modell_{ij}}$, $Modell_{ij} = \frac{Zeilensumme_i * Spaltensumme_j}{n}$

		Belohnung Leckerli	Belohnung Lob	Insg
Tanzen sie?	Ja	28	48	76
	Nein	10	114	124
	Insg	38	162	200

$\rightarrow$

		Belohnung Leckerli	Belohnung Lob
Tanzen sie?	Ja	$\frac{(28-14.44)^2}{14.44} = 12.73$	$\frac{(48-61.56)^2}{61.56} = 2.99$
	Nein	$\frac{(10-23.56)^2}{23.56} = 7.80$	$\frac{(114-100.44)^2}{100.44} = 1.83$

$\chi^2 = 12.73 + 2.99 + 7.80 + 1.83 = \underline{25.35}, df = 1$
$\chi^2 = 25.35 > \chi^2_{kr99\%}(df = 1) = 6.63 > \chi^2_{kr95\%}(df = 1) = 3.84$
$\rightarrow H_0$ wird verworfen $\rightarrow$ Es besteht ein signifikanter Zusammenhang mit $\alpha = 0.05\%$ *und* $\alpha = 0.01\%$.

Yates Korrektur

- Bei 2x2 Tabellen tendiert χ^2 Test zu Typ 1 Fehlern False Positive
- χ^2 zu groß
- $\chi^2 = \sum \frac{(|Beobachtet_{ij} - Model_{ij}| - 0.5)^2}{Modell_{ij}}$
- Reduziert wohl zu viel
  Howell, D.C. (2006): *Statistical methods for psychology*

"Although it is worth knowing about, it's probably best ignored" $_{AndyField}$

Annahmen

Unabhängigkeit der Zellen

- Jeder Proband darf nur zu einer Zelle zählen
- $\rightarrow$ **Nicht anwendbar für abhängiges Design!** Erst Leckerli, dann Lob

Zu erwartende Werte ($Modell_{ij}$) größer als 5 für jede Zelle

- 20% $<$ 5 Tolerierbar aber hoher Anstieg der Typ 2 Fehler (False Negative, Effekt übersehen)
- $Modell_{ij} < 1$ nicht tolerierbar
- Genauer: Howell, D.C. (2006): *Statistical methods for psychology*
- Größere Stichprobe oder Fishers exakter Test kann hier helfen

Fishers exakter Test

Fisher, R.A. (1922): *On the interpretation of chi square from contingency tables, and the calculation of P*

- Im Grunde χ^2 Test mit exakt berechnetem p
- Gebaut für kleinen Stichproben
- Bei großen Stichproben unnötig und rechenintensiv

(Maximum) Likelihood Ratio

Grundidee:

- Berechne Modell mit maximierter Wahrscheinlichkeit des Auftretens der Daten
- Vergleiche Modell mit der Wahrscheinlichkeit die Daten zufällig unter H_0 zu sehen

Berechnung

- $L\chi^2 = 2 * \sum Beobachtet_{ij} * \ln \frac{Beobachtet_{ij}}{Modell_{ij}}$
- Interpretation wie χ^2-Test
- $L\chi^2 = 24.94$ für unser Beispiel

Bewertung

- Bei großen Stichproben kaum Unterschied zu χ^2-Test, bei kleinen Stichproben ist Likelihood Ratio sicherer

Effektstärke

- Odds Ratio Siehe Vorlesung zu Logistischer Regression
- $oddsratio = \frac{odds_{tanzen\ nach\ leckerli}}{odds_{tanzen\ nach\ lob}}$
- $odds_{tanzen\ nach\ leckerli} = \frac{leckerli\ und\ tanzen}{leckerli\ und\ nicht\ tanzen} = \frac{28}{10} = 2.8$
- $odds_{tanzen\ nach\ lob} = \frac{lob\ und\ tanzen}{lob\ und\ nicht\ tanzen} = \frac{48}{114} = 0.421$
- $oddsratio = \frac{2.8}{0.421} = 6.65$
- "Die Chance, dass die Katze nach den Leckerlis tanzt, ist 6.65 mal höher als nach Lob."
- wird in *R* mit Konfidenzintervallen geliefert wenn fisher = true
  - $\rightarrow$ Konfidenzintervalle sollten 1-Grenze nicht überschreiten

Standardisierte Residuen

- Residuum: Abweichung von beobachtetem Wert zum Modellwert
- $Residuum_{ij} = Beobachtet_{ij} - Modell_{ij}$
- Standardisiertes Residuum: normalisiertes (vergleichbares) Residuum
  $stdresiduum_{ij} = \frac{Beobachtet_{ij} - Modell_{ij}}{\sqrt{Modell_{ij}}}$
- Beachte die Ähnlichkeit zu χ^2, wir addieren nur nicht auf, also quadrieren wir auch nicht
- Standardisierte Residuen sind z-Scores für einzelne Werte

  $\rightarrow$ Item-spezifische Signifikanzwerte und Wahrscheinlichkeiten ableitbar
  $\rightarrow$ $\pm 1.96 \rightarrow$ Signifikant mit 95%
  $\rightarrow$ $\pm 2.58 \rightarrow$ Signifikant mit 99%

χ^2 Test in *R*

Gegeben Kontingenztabelle

```
library(gmodels)

leckerli <- c(10, 28)
lob <- c(114, 48)
katzentabelle <- cbind(leckerli, lob)

CrossTable(katzentabelle, fisher = TRUE, chisq = TRUE,
    expected = TRUE, sresid = TRUE, format = "SPSS")
```

χ^2 Test in *R* Output Kontingenztabelle

```
Total Observations in Table:  200
             |  leckerli | lob        |  Row Total |
-------------|-----------|------------|------------|
        [1,] |       10  |       114  |       124  | //Anzahl
             |   23.560  |   100.440  |            | //Zu erwartende Werte
             |    7.804  |     1.831  |            | //Chi-Square Anteil
             |    8.065% |    91.935% |    62.000% | //Prozent(Zeile)
             |   26.316% |    70.370% |            | //Prozent(Spalte)
             |    5.000% |    57.000% |            | //Prozent(Insgesamt)
             |   -2.794  |     1.353  |            | //Std. Residuen
-------------|-----------|------------|------------|
        [2,] |       28  |        48  |        76  | //Std. Residuen zeigen
             |   14.440  |    61.560  |            | //signifikanten Unterschied
             |   12.734  |     2.987  |            | //bei leckerli (95% und 99%),
             |   36.842% |    63.158% |    38.000% | //aber keinen signifikanten
             |   73.684% |    29.630% |            | //Unterschied bei lob.
             |   14.000% |    24.000% |            |
             |    3.568  |    -1.728  |            |
-------------|-----------|------------|------------|
Column Total |       38  |       162  |       200  |
             |   19.000% |    81.000% |            |
-------------|-----------|------------|------------|
```

χ^2 Test in *R* Output Signifikanztests

```
Pearson's Chi-squared test
Chi^2 =  25.35569     d.f. =  1     p =  4.767434e-07          //Hochsignifikant

Pearson's Chi-squared test with Yates' continuity correction       //Ignorierbar
Chi^2 =  23.52028     d.f. =  1     p =  1.236041e-06

Fisher's Exact Test for Count Data                          //Fishers Exakter Test
Sample estimate odds ratio:  0.1519927

Alternative hypothesis: true odds ratio is not equal to 1     //Fisher Two sided
p =  1.311709e-06
95% confidence interval:  0.06086544 0.352389

Alternative hypothesis: true odds ratio is less than 1      //Fisher One sided A
p =  7.7122e-07
95% confidence interval:  0 0.3131634

Alternative hypothesis: true odds ratio is greater than 1   //Fisher One sided B
p =  0.9999999
95% confidence interval:  0.07015399 Inf

       Minimum expected frequency: 14.44                    //Sollte größer 5 sein
```

Log-Lineares Modell

– Mehr als 2 Variablen

Beispiel : Können wir Katzen und Hunden tanzen beibringen?

χ^2 als lineare Regression

Kombiniere:

- Regressionsformel $\hat{Y}_i = (b_0 + b_1 * X) + \varepsilon_i$
- Datentabelle:

		Belohnung Leckerli	Belohnung Lob	Insg
Tanzen sie?	Ja	28	48	76
	Nein	10	114	124
	Insg	38	162	200

$\rightarrow$ Coding-Tabelle für die tanzenden Katzen

Dummy(Belohnung)	Dummy(Tanzen)	Interaktion	Häufigkeit
0	0	0	28
0	1	0	10
1	0	0	48
1	1	1	114

Log-Lineares Modell

Coding-Tabelle für die tanzenden Katzen

Dummy(Belohnung)	Dummy(Tanzen)	Interaktion	Häufigkeit
0	0	0	28
0	1	0	10
1	0	0	48
1	1	1	114

Lineares Modell:

– $outcome = b_0 + b_1 * Belohnung + b_2 * Tanzen + b_3 * Interaktion + \varepsilon_i$

Logarithmus macht kategorische Verteilung linear

– $ln(O_i) = ln(Modell) + ln(Fehler)$

– $ln(O_{ij}) = b_0 + b_1 * Belohnung + b_2 * Tanzen + b_3 * Interaktion + ln(\varepsilon_i)$

Log-Lineares Modell

Dummy(Belohnung)	Dummy(Tanzen)	Interaktion	Häufigkeit
0	0	0	28
0	1	0	10
1	0	0	48
1	1	1	114

- $ln(O_{ij}) = b_0 + b_1 * Belohnung + b_2 * Tanzen + b_3 * Interaktion + ln(\varepsilon_i)$

- $ln(O_{Leckerli,Ja}) = b_0 + 0 + 0 + 0 \rightarrow ln(28) = b_0 = 3.332$
- $ln(O_{Lob,Ja}) = b_0 + b_1 + 0 + 0 \rightarrow b_1 = ln(48) - 3.332 = 0.539$
- $ln(O_{Leckerli,Nein}) = b_0 + 0 + b_2 + 0 \rightarrow b_2 = ln(10) - 3.332 = -1.029$
- $ln(O_{Lob,Nein}) = b_0 + b_1 + b_2 + b_3 \rightarrow b_3 = ln(114) - 3.332 - 0.539 + 1.029 = 1.894$

Log-Lineares Modell:

- $ln(O_{ij}) = 3.332 + 0.539 * Belohnung - 1.029 * Tanzen + 1.894 * Interaktion + ln(\varepsilon_i)$

Log-Lineares Modell

Log-Lineares Modell:

- $ln(O_{ij}) = 3.332 + 0.539 * Belohnung - 1.029 * Tanzen + 1.894 * Interaktion + ln(\varepsilon_i)$
- Wenn man die Interaktion weglässt, erhält man χ^2 als lineares Modell $ln(O_i) = 2.67 + 1.45 Belohnung + 0.49 Tanzen + ln(\varepsilon_i)$
- Generell sind t-Test, ANOVA und χ^2 alle analog in lineare Modelle übersetzbar

mit 3 Variablen A, B und C:

- $ln(O_{ijk}) = b_0 + b_1 * A_i + b_2 * B_j + b_3 * C_k + b_4 * AxB_{ij} + b_5 * AxC_{ik} + b_6 * BxC_{jk} + b_7 * AxBxC_{ijk} + ln(\varepsilon_i)$

Genauer/Mathematischer: Tabachnick,B.G.& Fidell, L.S. (2007): *Using multivariate statistics*

Fitness Log-Lineares Modell

– ε nahezu 0 wegen Interaktionstermen

$\rightarrow$ Gesamte Variation wird vom Modell erklärt

$\rightarrow$ **Gesättigtes Modell** Saturated

Fitnessoptimierung mit hierarchischem Entfernen der Variablen

$\rightarrow$ Berechne Abweichung zwischen Vorhersage und Beobachtung

$\rightarrow$ Lösche komplexeste Interaktion solange sich die Likelihood Ratio nicht ändert

$\rightarrow$ Zuerst $AxBxC$ dann AxB, AxC, BxC dann A, B, C

$\rightarrow$ Stoppe, sobald Likelihood Ratio sich signifikant ändert

Annahmen

Unabhängigkeit der Zellen

- Jeder Proband darf nur zu einer Zelle zählen

Zu erwartende Werte ($Modell_{ij}$) größer als 5 für jede Zelle

- 20% < 5 Tolerierbar
- $Modell_{ij} < 1$ nicht tolerierbar
- Bei Problemen:

  $\rightarrow$ wenig einflussreiche Variablen eliminieren

    $\rightarrow$ Nicht signifikant bei höchster Interaktion und
    $\rightarrow$ Nicht signifikant bei wenigstens 1 mittlerer Interaktion

  $\rightarrow$ Variablenwerte zusammenfassen

    $\rightarrow$ *rot*,*gelb*,*grau* $\rightarrow$ *farbig*, *grau*

  $\rightarrow$ Mehr Daten
  $\rightarrow$ Akzeptanz

Effektstärke

- Zerlege Daten in Subsets aus 2 Variablen (*Katzen* und *Hunde*)
- Berechne Odds-Ratio $\rightarrow$ Siehe χ^2 Test

Log-Lineare Analyse in *R* Datenexploration

```
catsDogs<-read.delim("CatsandDogs.dat", header = TRUE)
catsDogs

table(catsDogs$Animal, catsDogs$Training, catsDogs$Dance)
xtabs(~Animal + Training + Dance, data = catsDogs)
```

```
, ,  = No
      Affection as Reward Food as Reward
  Cat                 114             10
  Dog                   7             14

, ,  = Yes
      Affection as Reward Food as Reward
  Cat                  48             28
  Dog                  29             20
```

Log-Lineare Analyse in *R* Datenexploration

```
library(gmodels)

justCats = subset(catsDogs, Animal=="Cat") //CrossTable kann nur mit 2
justDogs = subset(catsDogs, Animal=="Dog") //Variablen umgehen

CrossTable(justCats$Training, justCats$Dance, sresid = TRUE, prop.t=FALSE,
    prop.c=FALSE, prop.chisq=FALSE, format = "SPSS")
CrossTable(justDogs$Training, justDogs$Dance, sresid = TRUE, prop.t=FALSE,
    prop.c=FALSE, prop.chisq=FALSE, format = "SPSS")

Total Observations in Table:  70
                   | justDogs$Dance
  justDogs$Training |        No |       Yes | Row Total |
-------------------|-----------|-----------|-----------|
Affection as Reward |         7 |        29 |        36 |  //Anzahl
                   |   19.444% |   80.556% |   51.429% |  //Prozent(Zeilen)
                   |    -1.156 |     0.757 |           |  //Std. Residuen
-------------------|-----------|-----------|-----------|
     Food as Reward |        14 |        20 |        34 |
                   |   41.176% |   58.824% |   48.571% |
                   |     1.190 |    -0.779 |           |
-------------------|-----------|-----------|-----------|  //Für Katzen
       Column Total |        21 |        49 |        70 |  //Siehe vorher
-------------------|-----------|-----------|-----------|
```

Log-Lineare Analyse in R χ^2 als LLM

```
catTable<-xtabs(~ Training + Dance, data = justCats)
catSaturated<-loglm(~ Training + Dance + Training:Dance,data = catTable,fit = TRUE)
summary(catSaturated)                                        //Gesättigtes Modell

Formula:
~Training + Dance + Training:Dance

...                                                                  //unwichtig

Statistics:
                 X^2 df P(> X^2)
Likelihood Ratio   0  0        1                          //Perfekte Vorhersage
Pearson            0  0        1
```

Log-Lineare Analyse in R χ^2 als LLM

```
catTable<-xtabs(~ Training + Dance, data = justCats)
catNoInteraction<-loglm(~ Training + Dance, data = catTable, fit = TRUE)
summary(catNoInteraction)                                //Ungesättigtes Modell
                                    //Fit=True berechnet zu erwartende Werte
Formula:
~Training + Dance
attr(,"variables")

...                                                                //unwichtig

Statistics:
                      X^2 df     P(> X^2)                  //=Chi^2 von vorher
Likelihood Ratio 24.93159  1 5.940113e-07                //Ganz schlechter Fit
Pearson          25.35569  1 4.767434e-07    //Modell signifikant anders als Daten
```

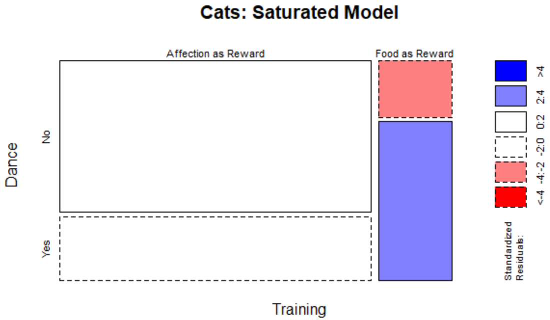
Log-Lineare Analyse in R

Wir stellen fest:

- Gesättigtes Modell = Perfekter Fit
- Entfernung des höchststufige Variable (FoodxAffection) erzeugt signifikante Abweichung
- $\rightarrow$ Wir rechnen mit gesättigtem Modell weiter

Log-Lineare Analyse in *R* Mosaic Plot

```
mosaicplot(catSaturated$fit, shade = TRUE, main = "Cats: Saturated Model")
```



- Standardisierte Residuen $> \pm 1.96 \rightarrow$ signifikant mit 95%
- Eingefärbt $\rightarrow$ Signifikant
- Linie gestrichelt $\rightarrow$ Weniger als erwartet
- Linie durchgezogen $\rightarrow$ Mehr als erwartet

Log-Lineare Analyse in *R*

Wir erhöhen die Variablenzahl und arbeiten mit 3 Interaktionsstufen

- Stufe 1: Training + Dance + Animal
- Stufe 2:
  - $\rightarrow$ Training x Dance
  - $\rightarrow$ Training x Animal
  - $\rightarrow$ Dance x Animal
- Stufe 3: Training x Dance x Animal

Log-Lineare Analyse in *R*

Schritt 1: Gesättigtes Modell erstellen

```
CatDogContingencyTable<-xtabs(~ Animal + Training + Dance, data = catsDogs)
caturated<-loglm(~ Animal*Training*Dance, data = CatDogContingencyTable)
summary(caturated)                          //Animal*Training*Dance = Abkürzung für
                                                     //alle möglichen Interaktionen
Formula:
~Animal * Training * Dance

...                                                                      //unwichtig

Statistics:
                 X^2 df P(> X^2)
Likelihood Ratio   0  0        1                                //Perfekte Vorhersage
Pearson            0  0        1
```

Log-Lineare Analyse in *R*

Schritt 2: Parsimony anstreben (Höchste Interaktion entfernen)

```
threeWay <- loglm(~ Animal + Training + Dance + Animal:Training +
    Animal:Dance + Dance:Training, data = CatDogContingencyTable)

//oder

threeWay<-update(caturated, .~. -Animal:Training:Dance)

summary(threeWay)

Formula:
. ~ Animal + Training + Dance + Animal:Training + Animal:Dance + Training:Dance

...                                                          //unwichtig

Statistics:
                      X^2 df     P(> X^2)
Likelihood Ratio 20.30491  1 6.603088e-06
Pearson          20.77759  1 5.158318e-06
```

Log-Lineare Analyse in *R*

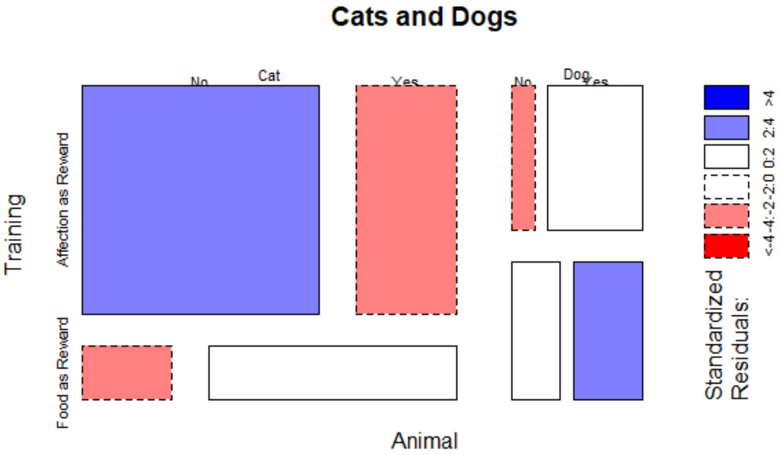
Schritt 3: Untersuche Differenz zwischen beiden Modellen

```
anova(caturated, threeWay)                              //Wir sind rechenfaul



LR tests for hierarchical log-linear models

Model 1:
 . ~ Animal + Training + Dance + Animal:Training + Animal:Dance + Training:Dance
Model 2:
 ~Animal * Training * Dance

          Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   20.30491  1
Model 2    0.00000  0   20.30491         1         1e-05
Saturated  0.00000  0    0.00000         0         1e+00          //Signifikant
```

Da der Unterschied signifikant ist, ist die Interaktion *Training x Dance x Animal* signifikant und darf nicht entfernt werden. Parsimony ist erreicht.
→ **STOP!!!**
Falls nicht signifikant, mache weiter mit Iteraktionen 2. Stufe, usw.

Log-Lineare Analyse in *R* Mosaic Plot

```
mosaicplot(CatDogContingencyTable, shade = TRUE, main = "Cats and Dogs")
```



Zusammenfassung

- 2 Kategorische Variablen
  - $\rightarrow$ χ^2 Test
  - $\rightarrow$ Bei kleiner Stichprobe Fishers exakter Test
  - $\rightarrow$ Yates Korrektur nett aber ignorierbar
  - $\rightarrow$ Alternativ Maximum Likelihood Ratio
  - $\rightarrow$ Odds-Ratio als Effektstärke
  - $\rightarrow$ Standardisierte Residuen als Signifikanztest der Zellen
- Mehr als 2 Kategorische Variablen
  - $\rightarrow$ Loglineare Analyse
  - $\rightarrow$ Starte mit gesättigtem Modell und erzeuge hierarchisch Parsimony
  - $\rightarrow$ Mosaic-Plots zeigen Verteilung sowie Standardisierte Residuen (Signifikanz)
  - $\rightarrow$ Odds-Ratio auf Subsets als Effektstärke