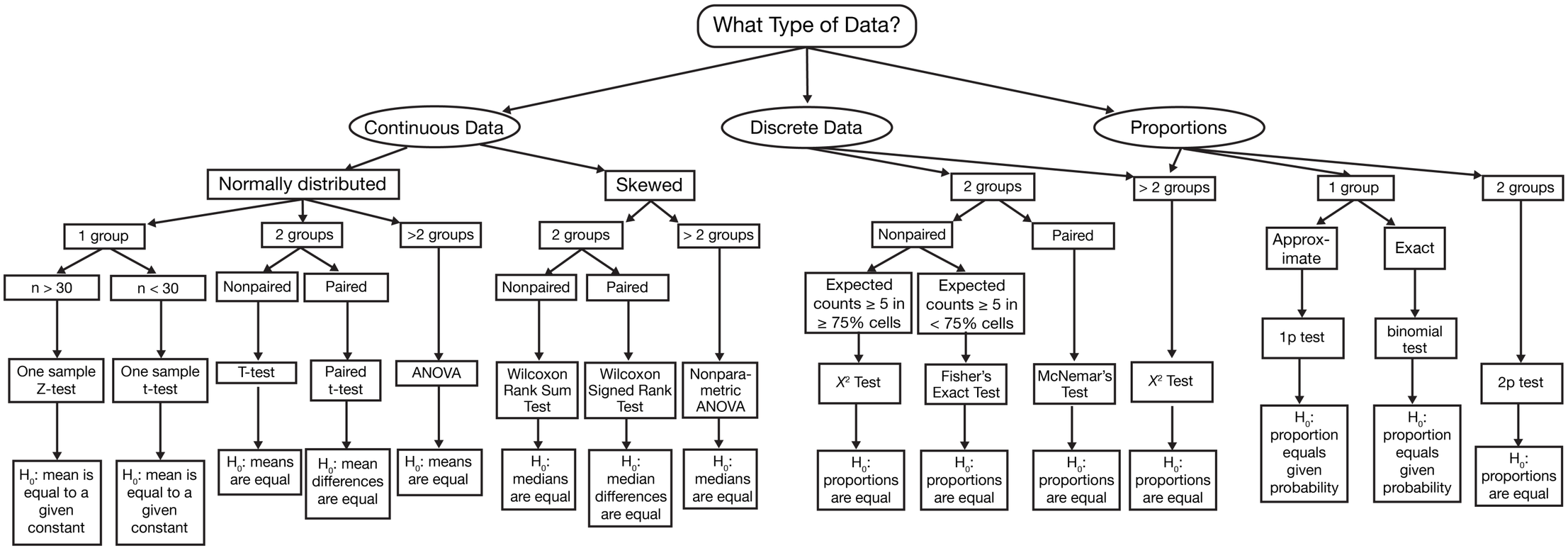
Anwendungsorientierte Analyseverfahren

Das GLM

Prof. Dr. Michael Scharkow

Sommersemester 2024

Klassische Statistiklehre



Quelle: https://onishlab.colostate.edu/wp-content/uploads/2019/07/which_test_flowchart.png

DATENANALYSE ALS REZEPTSAMMLUNG

- In der klassischen Statistikausbildung (auch bei uns) als Rezeptesammlung:
  - Mittelwerte in (genau) zwei Gruppen vergleichen - T-Test
  - Mittelwerte in mehr als zwei Gruppen vergleichen - Varianzanalyse (ANOVA)
  - Zusammenhänge von kategoriellen Variablen testen - χ^2-Test
  - …
- Fokus auf Unterschieden und Spezifika statt auf Gemeinsamkeiten
- Viele Verfahren sind aber mindestens funktional, oft auch mathematisch äquivalent!

BEISPIELSTUDIE: AUTY & LEWIS (2004)

> There has been little attempt to understand the influence on children of branded products that appear in television programs and movies. A study exposed children of two different age groups (6–7 and 11–12) in classrooms to a brief film clip. Half of each class was shown a scene from Home Alone that shows Pepsi Cola being spilled during a meal. The other half was shown a similar clip from Home Alone but without branded products. All children were invited to help themselves from a choice of Pepsi or Coke at the outset of the individual interviews.

BEISPIELSTUDIE: DATEN

id	pepsi_placement	pepsi_chosen
14	1	1
51	1	0
80	0	0
90	0	0
92	0	0

BEISPIELSTUDIE: CHI-QUADRAT TEST

Kreuztabelle (Spaltenprozente)

pepsi_chosen	no_placement	placement
0	57	37
1	43	63

Chi-Quadrat Test

Chi2(1)	p	Cramer's V (adj.)	Cramers_v_adjusted CI
4.14	0.042	0.17	(0.00, 1.00)

BEISPIELSTUDIE: BIVARIATE KORRELATION

Pearson Korrelation

Parameter1	Parameter2	r	95% CI	p
pepsi_placement	pepsi_chosen	0.20	(0.01, 0.38)	0.042

Alternative hypothesis: true correlation is not equal to 0

Kendall Korrelation

Parameter1	Parameter2	tau	z	p
pepsi_placement	pepsi_chosen	0.20	2.03	0.043

Alternative hypothesis: true tau is not equal to 0

BEISPIELSTUDIE: MITTELWERTVERGLEICHE

t-Test

Difference	95% CI	t(103)	p	d
-0.20	(-0.39, -0.01)	-2.06	0.042	-0.41

ANOVA

Parameter	Sum_Squares	df	Mean_Square	F	p	Eta2
pepsi_placement	1.03	1	1.03	4.23	0.042	0.04
Residuals	25.10	103	0.24			

GEMEINSAMKEITEN UND UNTERSCHIEDE DER VERFAHREN

- dieselbe Testentscheidung (signifikanter Unterschied zwischen den Gruppen bzw. signifikanter Zusammenhang zwischen Placement und Produktwahl).
- bei 3 Verfahren exakt gleichen p-Wert (d.h. die dahinterliegende Berechnung ist identisch), beim Chi-Quadrat-Test einen (leicht) abweichenden (d.h. Berechnung ist nicht identisch).
- die Verfahren unterscheiden sich vor allem im Modelloutput
- manchmal nur globale Teststatistiken (Chi-Quadrat, F-, t-Wert), manchmal auch Konfidenzintervalle oder Effektgrößen
- auch wenn es z.T. substanziell-statistische Unterschiede gibt, unterscheiden sich vor allem die Konventionen des Berichtens

Das Allgemeine Lineare Modell (General linear model, GLM)

“The only formula you’ll ever need.” Andy Field

DATENANALYSE ALS STATISTISCHE MODELLIERUNG

- Datenanalyse als Anwendung und Test bestimmter statistischer **Modelle**
- ein statistisches Modell ist eine vereinfachte Vorstellung, wie die beobachteten Daten zustande kommen (könnten)
- wir wenden diese Modell an und prüfen, wie gut die empirischen Daten dazu passen

$$\text{outcome}_i = \text{Model}_i + \text{error}_i$$

- beobachtete Daten (Outcome) als Summe von modellierten und nicht-modellierten Zusammenhängen

DAS NULLMODELL

Frage: Wenn wir nur einen Schätzwert a für Y haben, welcher ist der beste Schätzer?

$$Y_i = a + \epsilon_i$$

- das beste a ist dasjenige, das den Fehler ϵ minimiert ($\epsilon_i = Y_i - a$)
- bester Schätzer = kleinste Summe quadrierter Abweichungen ϵ_i von y
- Kriterium der *least squares* -> Ordinary Least Squares (OLS)

Antwort: Mittelwert x als der beste Modellkoeffizient im Nullmodell

Problem: damit erklärt das Modell aber nichts, es fehlt eine Prädiktorvariable X

MODELLFORMEL FÜR DAS GLM (BIVARIAT)

$$Y_i = b_0 + b_1 X_i + \epsilon_i$$

- Grundidee, eine Variable Y (abhängige Variable, Outcome) durch ein statistisches Modell mit einem oder mehr Parametern b vorhersagen zu lassen
- Annahme: linearer Zusammenhang, d.h. Y hängt nur von b_0 und der durch b_1 gewichteten (unabhängige) Prädiktorvariable X ab
- b_0 = Intercept = Achsenabschnitt = vorhergesagter Wert von Y, wenn $X = 0$
- grundlegende Interpretation:
  - "je mehr X, desto mehr Y", wenn $b_1 > 0$, und
  - "je mehr X, desto weniger Y", wenn $b_1 < 0$.
- es bleibt ein Vorhersage- bzw. Residualfehler ϵ (der minimiert wird)

MODELLFORMEL FÜR DAS GLM (MULTIVARIAT)

$$Y_i = b_0 + b_1 X_1 + +b_2 X_2 + b_3 X_3 + \ldots + \epsilon_i$$

- weil der Modell eine lineare Gleichung ist, können wir problemlos mehrere Prädikorvariablen X hinzufügen
- Outcome Y als eine (gewichtete) Linearkombination der Prädiktorvariablen X_1 ... X_k
- Parameter $b_1, b_2, b3$ ... sind die Gewichte, mit denen die Prädiktoren X zur Vorhersage von Y beitragen
- Interpretation von jedem b ist dieselbe wie im bivariaten Fall
- Intercept b_0 ist der vorhergesagte Wert von Y, wenn **alle** $X_1 = X_2 = X_3 = 0$.

ANWENDUNGSFÄLLE DES GLM

- Wenn die Prädiktorvariablen X kategoriell sind, entspricht das GLM dem T-Test bzw. der Varianzanalyse.
- Wenn die Prädiktorvariablen X metrisch sind, entspricht das GLM der linearen Regression bzw. Korrelation.
- Man kann problemlos beliebig viele kategorielle und metrische Prädiktoren mischen.
- Die Interpretation ist immer dieselbe, d.h. man muss nur eine Interpretationsregel lernen.

ANNAHMEN UND ERWEITERUNGEN

- Annahme: Zusammenhang zwischen X und Y ist linear
  - wenn die Annahme nicht gerechtfertigt ist, kann man auch andere funktionale Zusammenhänge modellieren,
    siehe Sitzung zur logistischen Regression
- Annahme: Untersuchungseinheiten sind unabhängig voneinander
  - wenn die Annahme verletzt ist, kann man Abhängigkeiten zwischen Fällen modellieren,
    siehe Sitzung zu Multilevel-Modellen

WELCHE KENNZIFFERN SIND RELEVANT?

- Modellparameter bzw. Regressionskoeffizienten b geben die geschätzten Zusammenhänge bzw. Unterschiede wieder
- Koeffizienten haben einen Punktschätzer und einen Standardfehler bzw. ein Konfidenzintervall (Inferenzstatistik)
- (Null-)Hypothesentests der Koeffizienten = testen, ob die beobachteten Daten zur Nullhypothese $b = 0$ passen
- Modellgütemaße wie R^2 quantifizieren, wie gut das statistische Modell insgesamt die Werte von Y vorhersagen kann (Verhältnis von vorhergesagter und Residualvarianz)

MODELLVORHERSAGEN

- Regressionsmodelle sind Vorhersageinstrumente
- mit Hilfe der Regressionskoeffizienten kann man für jede Kombination von Prädiktoren X das Outcome Y vorhersagen
- Vorhersagen für einzelne Individuen oder spezifische Gruppen (siehe Sitzung Modellvorhersagen)
- vorhergesagte Werte für die Visualisierung von Unterschieden und Zusammenhängen verwenden
- Vorhersagen oft intuitiver zu verstehen als einzelne Parameterschätzungen

WIE IST NUN UNSER GLM-REZEPT?

1. Daten einlesen und Outcome Y deskriptiv auswerten
2. GLM spezifizieren (d.h. welches sind unsere Prädiktorvariablen?) und schätzen
3. Regressionskoeffizienten interpretieren (Vorzeichen, Größe, Konfidenzintervall, stat. Signifikanz)
4. Modellgüte und ggf. globale Teststatistik interpretieren
5. durch das Modell vorhergesagte Werte schätzen, vergleichen, visualisieren

BEISPIELSTUDIE: GLM

Modelloutput (Regressionstabelle)

Parameter	Coefficient	95% CI	t(103)	p	Std. Coef.	Fit
(Intercept)	0.43	(0.29, 0.57)	6.24	< .001	0.00	
pepsi placement	0.20	(0.01, 0.39)	2.06	0.042	0.20	
AICc						153.96
R2						0.04
R2 (adj.)						0.03
Sigma						0.49

INTERPRETATION

- Intercept b_0: in der Kontrollgruppe (kein Placement, $X = 0$) vorhergesagte Wahrscheinlichkeit von .43 für Pepsi
- Regressionskoeffizient b_1: bei Placement ($X = 1$) ist die vorhergesagte Wahrscheinlichkeit für Pepsi .20 **höher** als ohne Placement
- der Regressionskoeffizient b_1 ist stat. signifikant ($p < .05$), d.h. er deckt sich nicht mit der Nullhypothese, dass es keinen Unterschied gibt
- Modellvorhersage bei Pepsi-Placement: $0.43 + 0.20 * 1 = .63$ in der Placement-Bedingung
- R^2: das Modell kann 4% der Varianz im Outcome Y erklären, der Rest bleibt unerklärt.

WAS SIND DIE NACHTEILE DER GLM-PERSPEKTIVE?

- viele SozialwissenschaftlerInnen haben es anders gelernt und verinnerlicht (“Warum machst du nicht T-Test statt Regression?”).
- Fachzeitschriften und Reviewer haben bestimmte Erwartungen und Vorgaben, AutorInnen präsentieren daher t-Test, ANOVA, etc.
- für Lektürekompetenz müssen wir (leider!) weiterhin auch die anderen Verfahren interpretieren können

LITERATUR

Andy Field. The General Linear Model. https://www.youtube.com/watch?v=7cSArk7tU4w

Auty, S., & Lewis, C. (2004). Exploring children's choice: The reminder effect of product placement. *Psychology & Marketing*, 21(9), 697-713.

Fragen?

Praktische Übung

HINWEISE ZUR R-ÜBUNG

- Kapitel auf https://stats.ifp.uni-mainz.de/ba-aa-vl lesen und parallel RStudio öffnen
- immer zuerst prüfen, ob das RStudio-Projekt “BA-AA-VL” aktiv ist (oben rechts im Fenster)
- im Files-Reiter die R-Datei zur Sitzung öffnen, z.B. heute `glm.R`
- beim Lesen in der R-Datei Code blockweise ausführen (immer `STRG+ENTER` bzw. `CMD+ENTER`)
- Output ansehen, Code nachvollziehen, modifizieren