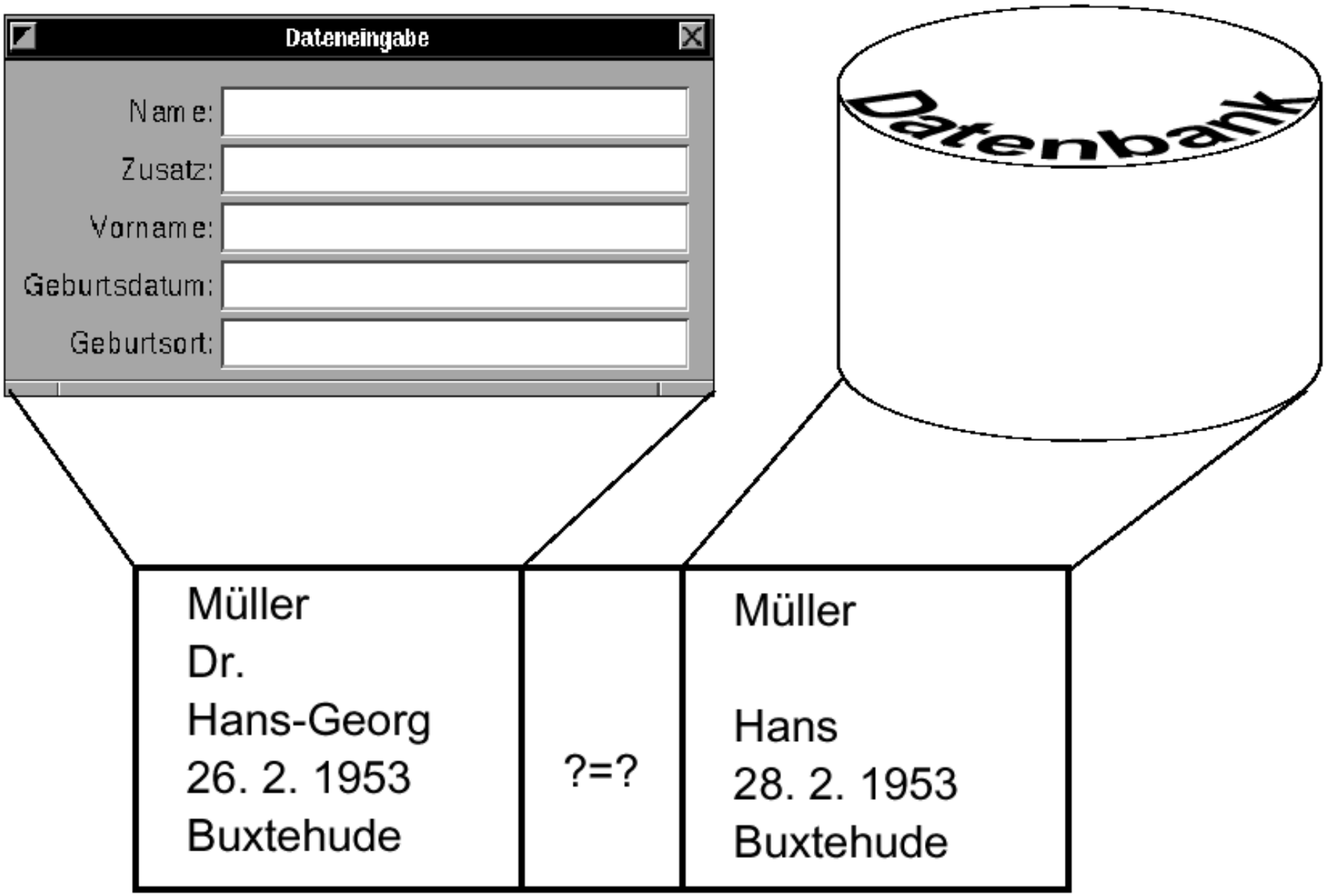
Stochastisches Matchen – Schätzung der Eingabefehlerraten

Klaus Pommerening
IMBEI
Oberseminar Medizin-Informatik
3. 6. 2003


Dateneingabe
Name:
Zusatz:
Vorname:
Geburtsdatum:
Geburtsort:
Datenbank
Müller
Dr.
Hans-Georg
26. 2. 1953
Buxtehude
?=?
Müller
Hans
28. 2. 1953
Buxtehude

I. Entscheidungen mit drei Ausgängen

Entscheidungssituation

$Y = M \cup U$ disjunkte Zerlegung

γ Beobachtungsfunktion

Γ Merkmalsraum

τ Entscheidungsfunktion

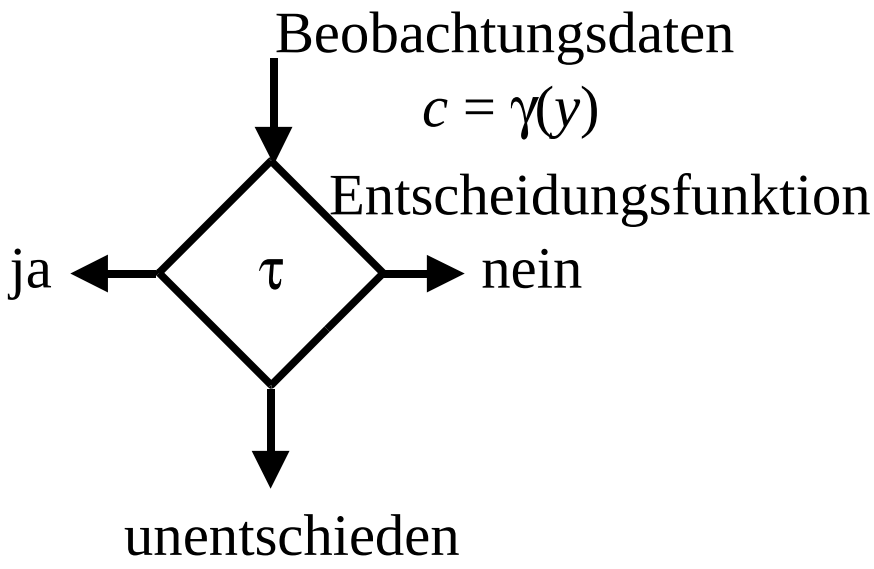
E = {ja, unentschieden, nein}

Alle Mengen endlich.

Aufgabe: Finde τ so, dass möglichst

- $\tau(\gamma(y))$ = ja, wenn $y \in M$,
- $\tau(\gamma(y))$ = nein, wenn $y \in U$.

[Entscheidung zwischen M und U aufgrund der Beobachtung $\gamma(y)$.]



Entscheidungsfehler

- **Fehler 1. Art** (falsches Ja): $y \in U$ mit $\tau(\gamma(y)) =$ ja.
  - Fehlerrate 1. Art: $\alpha(\tau) = P(\text{ja}|U) = \Sigma_{c\in\Gamma}\, P(c|U)\, P(\text{ja}|c)$
  - Gewichtsfaktor $u(c) = P(c|U)$.

- **Fehler 2. Art** (falsches Nein): $y \in M$ mit $\tau(\gamma(y)) =$ nein.
  - Fehlerrate 2. Art: $\beta(\tau) = P(\text{nein}|M) = \Sigma_{c\in\Gamma}\, P(c|M)\, P(\text{nein}|c)$
  - Gewichtsfaktor $m(c) = P(c|M)$.

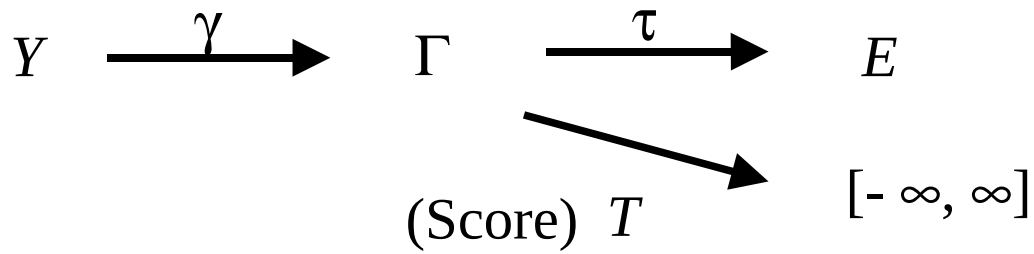
- **Unentschiedenheitsrate:**
  - $\eta(\tau) = P(\text{un.}) = \Sigma_{c\in\Gamma}\, P(c)\, P(\text{un.}|c)$
  - Gewichtsfaktor $n(c) = P(c)$.

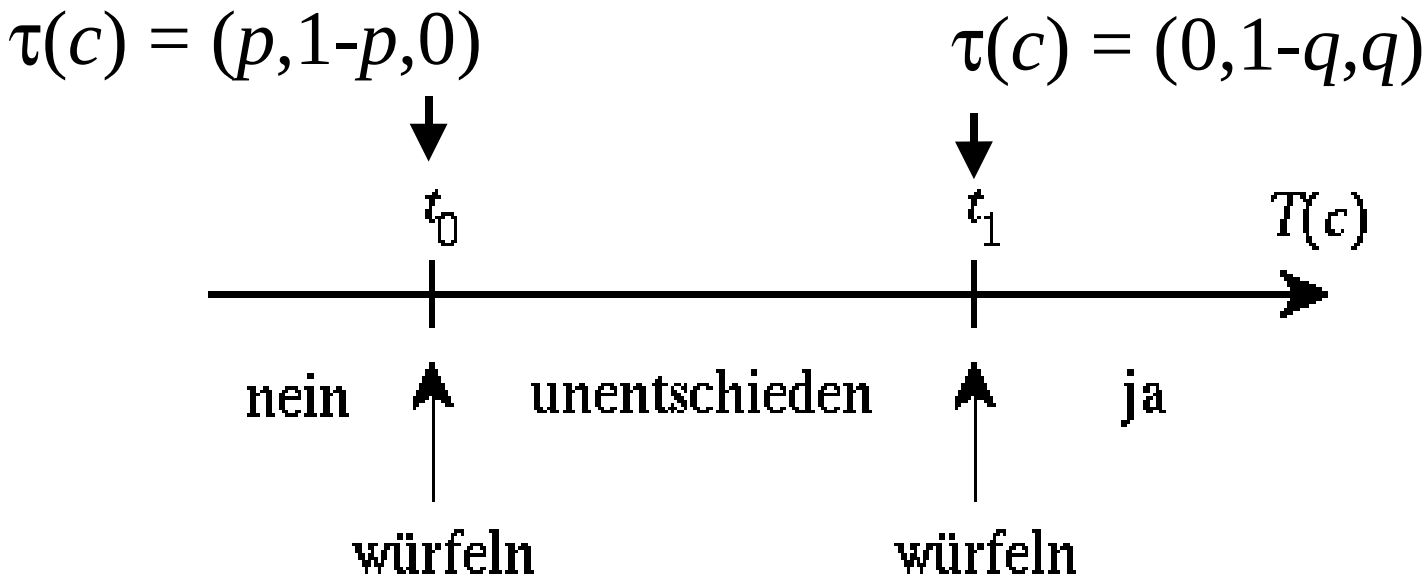
Optimierungsziel

- Finde eine Entscheidungsfunktion τ, die
  - bei gegebenen Schranken für die Fehlerraten 1. und 2. Art
  - die Unentschiedenheitsrate minimiert.
- Randomisierte Entscheidungsfunktion

$\tau: \Gamma \longrightarrow [0,1]^3$ mit $\tau_1 + \tau_2 + \tau_3 = 1$ konstant.

Schwellenwert-Verfahren





Die Gewichtsfunktion

Das (i. w.) optimale stochastische Entscheidungsverfahren ist das Schwellenwert-Verfahren zur **Gewichtsfunktion**

$$w := \log m/u : \Gamma \longrightarrow [-\infty, \infty]$$

D. h. $w(c) = \log \dfrac{P(c|M)}{P(c|U)}$ (»Likelihood Ratio«).

Falls $\Gamma = \Gamma_1 \times \cdots \times \Gamma_K$ (direktes Produkt, Gruppen unabhängiger Merkmale), wird die Gewichtsfunktion entsprechend additiv zerlegt.

II. Match-Entscheidungen (Abgleich, Record Linkage)

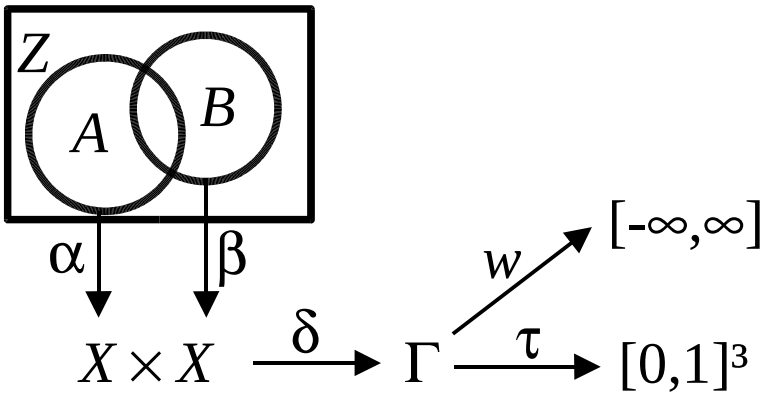
Das Match-Problem

- Grundmenge Z (gedachte Population).
- Teilmengen $A, B \subseteq Z$ (mögliche Überschneidung).
- $M = \{(a,b) \in A \times B \mid a = b\}$ (»Matches«),
- $U = \{(a,b) \in A \times B \mid a \neq b\}$ (»unpassende Paare«).
- $Y = A \times B = M \cup U$ (disjunkt).
- Datenerzeugende Prozesse (fehlerbehaftet)

  $$\alpha: A \to X, \qquad \beta: B \to X.$$

- Entscheidungsproblem = Match-Problem.

Die Match-Situation



Vergleichsfunktion

$\delta: X \times X \rightarrow \Gamma.$

(z. B. $\Gamma = X \times X$, δ = identische Abb. oder Γ = {gleich, ungleich}).

- Beobachtungsfunktion $\gamma = \delta_o(\alpha,\beta): A \times B \rightarrow \Gamma.$

  - **Homonymfehler** = Fehler 1. Art,
  - **Synonymfehler** = Fehler 2. Art.

Fehlerraten in den Daten

durch Eingabefehler + Änderungen.

- Datenerzeugender Prozess:
  $\alpha: Z \to X$.
- (Werteabhängige/lokale) Fehlerrate:
  $\varepsilon_\alpha: X \to [0,1]$, $\varepsilon_\alpha(x) = P(\alpha(z) \neq x \mid \xi(z) = x)$
  ($\xi(z)$ = wahrer Wert).
- Globale Fehlerrate:
  $\underline{\varepsilon}_\alpha = P(\alpha(z) \neq \xi(z))$.

Annahme beim Match-Problem:
Die Fehlerraten gelten auf A, B und $A \cap B$.

Merkmalshäufigkeiten

Annahmen:

- Die Ausprägung $x \in X$ kommt in Z, A, B und $A \cap B$ mit der Wahrscheinlichkeit $p(x)$ vor.
- Die Fehlerraten kompensieren sich so weit, dass auch

$$\begin{aligned} p(x) &= P(\xi(a) = x) \\ &\approx P(\alpha(a) = x) \\ &\approx P(\beta(b) = x). \end{aligned}$$

III. Gewichte beim Matchen

Bestimmung der Gewichte

(für jeweils unabhängige Merkmalsblöcke)

- Im allgemeinen nicht verfügbar, bestenfalls schätzbar.
- Zwei typische Fälle:
  - Prüfung auf Übereinstimmung: $\Gamma = \{\text{gleich, ungl.}\}$ (»globale Gewichte«).
  - Prüfung anhand der Werte: $\Gamma = X \times X$ (»werteabhängige/lokale Gewichte«).
- Newcombe-Faustregel: Gewichtsabweichung ≤ 0.3 irrelevant (Faktor 2 im LR).
  - konservativer: Gewichtsabweichung ≤ 0.1.

Globale Gewichte

Approximationsformeln:

$$w(\text{gleich}) \approx \log \frac{1}{\sum_{x \in X} p(x)^2}$$

$$w(\text{ungleich}) \approx \log \frac{\underline{\varepsilon}_\alpha + \underline{\varepsilon}_\beta}{1 - \sum_{x \in X} p(x)^2}$$

Annahmen dabei: Globale Fehlerraten klein ($\underline{\varepsilon}_\alpha, \underline{\varepsilon}_\beta \leq 5\%$).

Bemerkung: $P(\text{ungleich}|M) \approx \underline{\varepsilon}_\alpha + \underline{\varepsilon}_\beta$

$$P(\text{ungleich}|U) \approx 1 - \sum_{x \in X} p(x)^2$$

Anwendung: Gleichverteiltes Merkmal

X gleichverteilt mit *n* Ausprägungen, Annahme: Fehlerraten i. w. von Ausprägung unabhängig. Dann

$$w(\text{gleich}) \approx \log n,$$

$$w(\text{ungleich}) \approx \log \eta \quad [+ \Delta_X].$$

$\eta = \underline{\varepsilon}_\alpha + \underline{\varepsilon}_\beta$ Summe der globalen Fehlerraten.

Korrektursummand $\Delta_X = \log n/(n\text{-}1)$

$= 0.30, 0.18, 0.12, 0.10$ für $n = 2, 3, 4, 5$.

Unbekannt (d. h. zu schätzen) nur η

Beispiel: Geburtsmonat

Gleichverteiltes Merkmal mit n = 12.

Annahme: Beide globalen Fehlerraten ≈ 5%.

$$w(\text{gleich}) \approx \log 12 \approx 1.08,$$
$$w(\text{ungleich}) \approx \log 0.10 \approx -1.00.$$

[Bei Berücksichtung der Ungleichverteilung 1.03 bzw. –0.97.]

Werteabhängige (lokale) Gewichte

Approximationsformeln:

$$w(x,x) \approx -\log p(x),$$

$$w(x,x') \approx \log\left[\frac{\varepsilon_\alpha(x)}{p(x)} + \frac{\varepsilon_\beta(x')}{p(x')}\right] - \log(n-1) \quad \text{für } x \neq x'$$

Annahmen dabei: Werteabhängige Fehlerraten klein ($\varepsilon_\alpha, \varepsilon_\beta \leq 5\%$).

[**Achtung:** Auch im Ungleich-Fall i. d. R. wesentlich von globalen Gewichten verschieden.]

Beispiel: Geburtstag

$w(x,x) \approx -\log(12/365.5) \approx 1.48$ für $x = 1, ..., 28,$

...

$\approx -\log(7/365.5) \approx 1.72$ für $x = 31,$

$w(x,x') \approx -0.99$ für $x, x' = 1, ..., 28$ verschieden,

...

≈ -0.85 für $x = 30, x' = 31$

[bei Fehlerraten von konstant 5%].

Es lohnt sich, die Werteabhängigkeit zu berücksichtigen.

Zu schätzende Größen (allgemeiner Fall)

- Wahrscheinlichkeiten $p(x)$ [falls nicht vorgegeben]:
  - Schätzung als relative Häufigkeiten:
  - B = Datenbank „schon erfasster Fälle“,
  - N_B = Anzahl der Fälle in B,
  - $N_X(x)$ = Häufigkeit des Wertes x im Merkmal X,
  - $p(x) \approx N_X(x)/N_B$.
- Fehlerraten $\varepsilon_\alpha(x)$, $\varepsilon_\beta(x)$?
  - Komplizierter, zusätzliche Annahmen nötig.

Iterativer Abgleich

- Fälle laufen einzeln ein:
  - falls schon vorhanden: Match,
  - falls neu: Aufnahme in Datenbank.
- Realistisch für PID-Dienst, Register.
- Vorteil: Beim einzelnen Abgleich sind die „Erfahrungen“ der bisherigen Abgleiche verwendbar.
- Modell „Abgleich zweier Dateien“ noch geeignet?

IV. Schätzung der Fehlerraten

Modell I – gleiche Fehlerraten

Annahme für Modell I: Beide Fehlerraten ε_α, ε_β sind gleich, also beschrieben durch gemeinsame Funktion

$$\varepsilon: A \cup B \to [0,1].$$

Modell I – a) globale Fehlerraten

- Für globale Fehlerraten Mittelwert $\underline{\varepsilon}$ schätzen aus

  $$2\underline{\varepsilon} \approx P(\text{ungleich}|M) \approx \text{err}_X/N_M$$

  - N_M = Zahl der bisherigen Treffer (= erfolgreichen Matchvorgänge),
  - err_X = Zahl der Abweichungen in X dabei.
- Zähler mitführen, Startwert ε_0 plausibel vorgeben.

Modell I – b) werteabhängig

$$N_M(x) \approx p(x)\, N_M$$

$$\mathrm{err}_X(x) \approx \varepsilon(x)\, p(x)\, N_M$$

$$\varepsilon(x) \approx \frac{\mathrm{err}_X(x)}{N_M(x)} \approx \frac{\mathrm{err}_X(x)}{N_X(x)} \frac{N_B}{N_M}$$

Zähler $\mathrm{err}_X(x)$:

Bei Match mit $\alpha(a) = x$, $\beta(b) = x'$, $x \neq x'$
$\mathrm{err}_X(x)$ und $\mathrm{err}_X(x')$ je um ½ erhöhen.

Modell I – Problem

$\mathrm{err}_X(x)$ kommt bei seltenen Werten
[das sind fast alle]
kaum von der 0 weg.

Auch $\mathrm{err}_X(x) = 1$ führt zu sehr ungenauer Schätzung.

Besser handhabbar: Modell, das viele Werte poolt.

Modell II – sicher/unsicher

Annahme für Modell II:

Es gibt „sichere“ und „unsichere“ Daten (geringe oder höhere Fehlerrate).

$$A = A_s \cup A_u, \quad B = B_s \cup B_u \text{ (disjunkt).}$$

δ_s δ_u ε_s ε_u

jeweils konstante Fehlerrate

Bedingungen im Modell II

Vier verschiedene Situationen im Match-Fall:

- $\delta_s + \varepsilon_s \approx P(\text{ungleich} \mid A_s \cap B_s)$
- $\delta_u + \varepsilon_s \approx P(\text{ungleich} \mid A_u \cap B_s)$
- $\delta_s + \varepsilon_u \approx P(\text{ungleich} \mid A_s \cap B_u)$
- $\delta_u + \varepsilon_u \approx P(\text{ungleich} \mid A_u \cap B_u)$

Gleichungen abhängig, weitere Annahme nötig.

Modell II – Zusatzannahme

Weitgehend fehlerfreie Quelle für „sicher“ –

- $\delta_s = 0$: **neu eingegebener Fall fehlerfrei und aktuell** (... so gut wie),
- ε_s = Änderungswahrscheinlichkeit für Datenbankeintrag,
- δ_u = Fehler bei Neueingabe,
- ε_u = Fehler in Datenbank + Änderungswahrscheinlichkeit

Modell II – benötigte Zähler

- N^{ss} = Matches sicher/sicher,
  err^{ss} = dabei in X aufgetretene Abweichungen.
- N^{us} = Matches unsicher/sicher,
  err^{us} = dabei in X aufgetretene Abweichungen.
- N^{su} = Matches sicher/unsicher,
  err^{su} = dabei in X aufgetretene Abweichungen.

Probleme

- Abhängigkeit zwischen Merkmalen
  - „Schwache“ Abhängigkeit vernachlässigen? (z. B. Vorname / Geburtsjahr)
  - „Direkte“ Abhängigkeit durch „Blockbildung“ oder verschiedene Match-Läufe berücksichtigen oder differenzierte Vergleichsfunktion? (z. B. Name / phonetischer Code)
- Festlegung der Schwellenwerte.
- Umgang mit fehlenden Werten.
- ...