

# **ANALÝZA KATEGORIZOVANÝCH DAT V *SOCIOLOGII***

Tomáš Katrňák

Fakulta sociálních studií  
Masarykova univerzita  
Brno

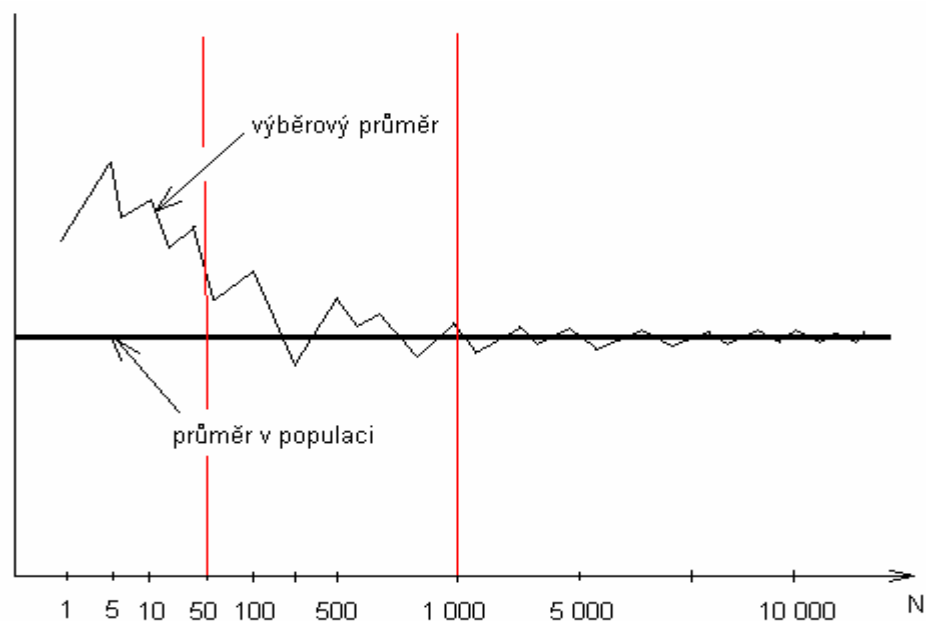
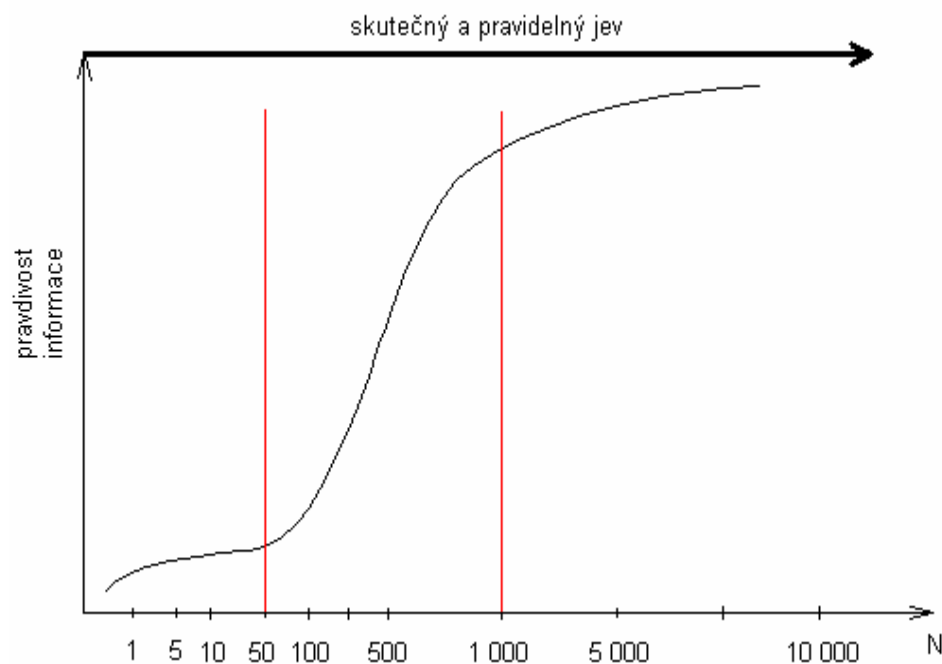
# Lekce 1: Úvod do analýzy kategorizovaných dat

## SOCIOLOGIE A STATISTIKA

- nadindividuální společenské struktury podmiňují lidské chování (Durkheim)
- společenské struktury lze pozorovat na základě statistik o lidském chování, pozorujeme novou skutečnost, z individuálního hlediska nerozpoznatelnou, ptací perspektiva, vymezuje a zároveň přináší informaci o tzv. **hromadném jevu**
  - **hromadný jev** je kolektivita nového řádu, její objevení souvisí s konstitucí moderní společnosti a ustavením sociologie a statistiky jako věd o sociálním životě v moderní společnosti
- **hromadný jev** je definován dostatečným počtem zkoumaných jednotek, protože až na základě určitého počtu (mnohosti) lze získat představu o pravidelnosti, struktuře a zákonitostech v sociálním životě (opakem je **individuální jev**)
  - kde vznikají sociální fakta, když nepřamení z psychiky člověka, ačkoliv jsou její nedílnou součástí? ptá se Durkheim
  - zdroje sociálních faktů leží v sociálních vazbách mezi lidmi, leží tedy v nadindividuálních sociálních strukturách, odpovídá Durkheim
- z tohoto důvodu sociologové pro pochopení sociálního života zkoumají nadindividuální sociální struktury, statistika a statistický aparát jim v tom pomáhají

## ZÁKON VELKÝCH ČÍSEL

- sociální jev je vždy hromadný jev, adjektivum sociální odkazuje k hromadnosti a sociálním vazbám (Simmel)
- všechny jevy (včetně sociálních) podléhají **zákonu velkých čísel** (jako první jej definoval francouzský matematik a statistik Poisson)
  - podle tohoto zákona se empirické údaje o jevu blíží skutečnosti s rostoucím počtem pozorovaných jednotek (když pozorujeme všechny jednotky, pozorujeme skutečnost), pravidelnost a pravá podstata jevu tedy vyvstává na povrch s rostoucím počtem pozorovaných případů



## PROMĚNNÉ A JEJICH DĚLENÍ

- podle slovního vyjádření hodnot proměnných:
  - **kvantitativní proměnné** (diskrétní & spojité)
  - **kvalitativní proměnné**
- podle vztahů mezi hodnotami jednotlivých proměnných:
  - **nominální** (název variant)
  - **ordinální** (název variant + uspořádání vertikální nebo horizontální)
  - **kardinální** (název variant + uspořádání + vzdálenost)
    - **intervalové** (o kolik je jedna hodnota větší než druhá),  $<-\infty; \infty>$ , neexistuje racionální 0 (např. teplota ve °C, 0 neznamena nepřítomnost teploty)
    - **poměrové** (kolikrát je jedna hodnota větší než druhá)  $<0; \infty>$ , 0 má racionální základ (např. věk, počet dětí, váha, životnost výrobku atd.)
- hranice mezi jednotlivými proměnnými nejsou neprůchodné, záleží na úhlu pohledu, např. členství v politické straně (nominální, ordinální) nebo vzdělání (nominální, ordinální, kardinální)
- proměnné vyššího řádu měření lze převést do nižšího řádu měření (tzv. ordinalizace nebo nominalizace proměnných)

## PROMĚNNÉ A JEJICH DĚLENÍ

- pod hlavičku kategorizované proměnné řadíme nominální, ordinální a kardinální poměrové proměnné
- kategorizované proměnné dělíme podle počtu variant:
  - **dichotomické** (binární, alternativní)
  - **polytomické** (vícekategoriální)
    - uspořádané kategorie (vertikálně, horizontálně)
    - neuspořádané kategorie (nominální proměnné)

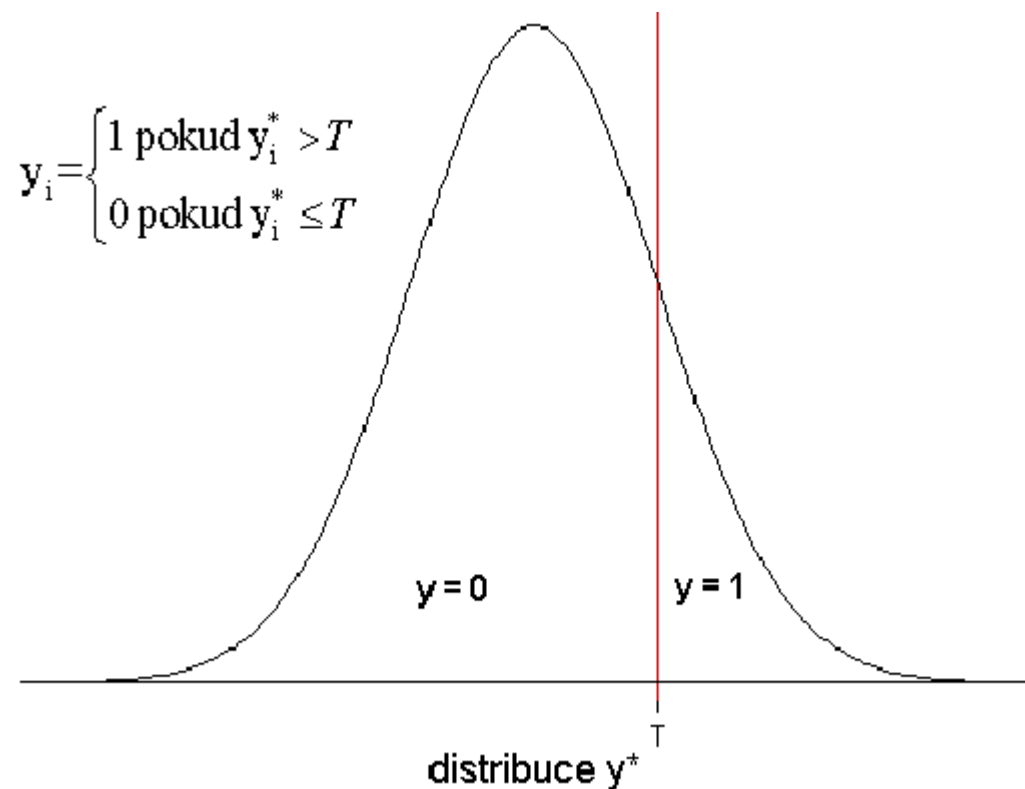
# TRANSFORMAČNÍ PŘÍSTUP VS. PŘÍSTUP LATENTNÍ PROMĚNNÉ

- dva přístupy v konceptualizaci kategorizovaných dat:
  - kategorizovaná data jsou inherentně **diskrétní proměnné** (nelineární pravděpodobnostní modely), statistika a biostatistika, **transformační přístup ke kategorizovaným datům**
    - výběrové varianty znaků a populační parametry se shodují, data modelujeme v měřené podobě,  $p$  (pravděpodobnost) závisle proměnné transformujeme do podoby lineární funkce nezávisle proměnných, tato funkce se nazývá **spojnice** (link), modely transformované do lineárních modelů spojnici se nazývají **zobecněné lineární modely (GLM)** ( $p$  je pak vyjádřena jako nelineární pravděpodobnostní model)
  - kategorizovaná data jsou nepozorované latentní proměnné (hovoří se o modelech latentních proměnných), tento přístup najdeme v ekonometrii a psychometrii, jedná se o tzv. přístup ke kategorizovaným datům jako k latentním proměnným
    - v populaci jsou kategorizovaná data kontinuální proměnné, pozorujeme je ovšem jako kategorizované (např. u proměnné dítě pozorujeme pouze dvě varianty, nicméně v populaci je to kontinuální proměnná, která když překročí určitou míru, tak pozorujeme její manifestaci), rozdíl mezi kontinuálními daty a kategorizovanými daty je v míře pozorovatelnosti, u kategorizovaných dat nevidíme jednotlivé hodnoty, ale pouze intervaly (proto se kategorizovaná data někdy nazývají jako omezené závisle proměnné), v populaci existují tendence, sklon a míra tolerance, přímo úměrné investicím a výnosům, pozorujeme pak jejich projevy skrze překročení míry tolerance

## TRANSFORMAČNÍ PŘÍSTUP VS. PŘÍSTUP LATENTNÍ PROMĚNNÉ

- debata o povaze kategorizovaných dat se táhne historií statistického uvažování (její počátek leží ve sporu K. Pearsona (latentní struktura) a G. U. Yulea (inherentní diskretnost) v první polovině 20. stol., dodnes tato debata není uzavřená, z obou dvou přístupů vycházejí odlišné numerické algoritmy k identifikaci modelů se závisle kategorizovanou proměnnou, jejich výsledky jsou nicméně totožné
- o kterých z následujících proměnných lze uvažovat jako o latentních?

- úspěch u zkoušky, souhlas s předmanželským sexuálním životem, pohlaví, participace na trhu práce, rodinný stav, přijetí na VŠ, sociální status, gravidita, četba časopisu Respekt, zaměstnanecká mobilita
- u latentní proměnné  $y^*$  předpokládáme, že je lineárně závislá na pozorované proměnné  $x$ , strukturálním vztahem vyjádřeno:
 
$$y^* = \mathbf{x}_i\beta + \varepsilon_i$$
 nebo pro jednoduchou proměnnou vyjádřeno vztahem
 
$$y^* = \alpha + \beta x_i + \varepsilon_i$$





## INDIVIDUÁLNÍ A AGREGOVANÁ DATA

- **individuální data**

- ukazují varianty znaků pro jednotlivá pozorování
- jednotlivé případy charakterizuje vždy jedna varianta zkoumané proměnné
- data jsou prezentována obvykle ve formě matice, v níž vždy jeden řádek odpovídá jednomu pozorování (případu) a jeden sloupec vždy jedné proměnné (znaku), pole matice pak zachycují varianty proměnných u jednotlivých pozorování (případů)

- **agregovaná data**

- ukazují počet opakujících se pozorování
- jednotlivé kombinace variant proměnných jsou charakterizovány počtem případů
- data jsou prezentována obvykle ve formě kontingenční tabulky, v řádcích a sloupcích tabulky jsou zkombinovány varianty proměnných, v polích tabulky jsou četnosti pozorování (počty případů) těchto variant

## AGREGOVANÁ DATA A JEJICH ANALÝZA POMOCÍ STATY

- **fully relational format of data** - každé pole tabulky odpovídá jednomu pozorování, pole tabulky jsou v matici soustředěné pod jednu proměnnou
- **folded (grouped) format of data** - pozorování je o polovinu méně než polí v tabulce, nicméně pozorování jsou soustředěná pod dvě proměnné (tedy do šířky matice)
- **příklad:**

|           | Age through 54 |         | Age through 55 and above |         |
|-----------|----------------|---------|--------------------------|---------|
|           | tolbutamine    | placebo | tolbutamine              | placebo |
| Dead      | 8              | 5       | 22                       | 16      |
| Surviving | 98             | 115     | 76                       | 79      |

(1) fully relational format

|    | agecat | exposed | died | pop |
|----|--------|---------|------|-----|
| 1. | 0      | 1       | 1    | 8   |
| 2. | 0      | 1       | 0    | 98  |
| 3. | 0      | 0       | 1    | 5   |
| 4. | 0      | 0       | 0    | 115 |
| 5. | 1      | 1       | 1    | 22  |
| 6. | 1      | 1       | 0    | 76  |
| 7. | 1      | 0       | 1    | 16  |
| 8. | 1      | 0       | 0    | 69  |

(2) folded format

|    | agecat | exposed | deaths | pop |
|----|--------|---------|--------|-----|
| 1. | 0      | 1       | 8      | 106 |
| 2. | 0      | 0       | 5      | 120 |
| 3. | 1      | 1       | 22     | 98  |
| 4. | 1      | 0       | 16     | 85  |

- podle typu dat volíme ve Statě syntax výpočtu, např. **logit** akceptuje (1), **blogit** akceptuje (2), **glogit** akceptuje (2), ale odhad není proveden jako ML, ale jako WLS, **glm** akceptuje jak (1), tak (2), obecně je ve Statě rozšířenější typ dat (1)

# Lekce 2: Analýza dvojrozměrných tabulek v sociologii

## LOGIKA A NOTACE KONTINGENČNÍCH TABULEK

- kontingenční tabulky jsou prvním (a nejstarším) krokem k analýze kategorizovaných dat
- např. kontingenční tabulka víra v posmrtný život podle pohlaví (zdroj: Agresti 1996:17)

| pohlaví | víra |         | pozorované četnosti: |          |          | očekávané četnosti: |          |          |                            |
|---------|------|---------|----------------------|----------|----------|---------------------|----------|----------|----------------------------|
|         | ano  | ne/neví |                      |          |          |                     |          |          |                            |
| žena    | 435  | 147     | $n_{11}$             | $n_{12}$ | $n_{1+}$ | $f_{11}$            | $f_{12}$ | $f_{1+}$ | $F_{11}$ $F_{12}$ $F_{1+}$ |
| muž     | 375  | 134     | $n_{21}$             | $n_{22}$ | $n_{2+}$ | $f_{21}$            | $f_{22}$ | $f_{2+}$ | $F_{21}$ $F_{22}$ $F_{2+}$ |
|         |      |         | $n_{+1}$             | $n_{+2}$ | $N$      | $f_{+1}$            | $f_{+2}$ | $f_{++}$ | $F_{+1}$ $F_{+2}$ $F_{++}$ |

- ve dvojrozměrné tabulce proměnná  $x$  má  $i$  úrovní (variant) a proměnná  $y$  má  $j$  úrovní (variant), pole v tabulce reprezentují  $ij$  možné výsledky, neboli velikost tabulky, taková tabulka se nazývá kontingenční tabulka (2 proměnné = dvojrozměrná, 3 proměnné = trojrozměrná, atd.), např. tabulka o rozměrech  $2 \times 2$  ( $i \times j$ ) má 4 pole (4 frekvence), tabulka o rozměrech  $3 \times 2 \times 2$  ( $i \times j \times k$ ) má 12 polí (12 frekvencí)
- $f_{ij}$  označuje pozorovanou (naměřenou) četnost v tabulce
- $F_{ij}$  označuje očekávanou (vypočítanou) četnost v tabulce za určitého předpokladu

## LOGIKA A NOTACE KONTINGENČNÍCH TABULEK

- každé  $f_{ij}$  v tabulce označuje počet (četnost) případů, které připadají na toto pole tabulky, neboli reprezentuje souběžný výskyt jednotlivých variant proměnných
- pomocí tabulkové notace ( $f_{ij}$ ) a frekvenčních vah [fweight= ] můžeme kontingenční tabulky vkládat do statistických programů a analyzovat je
- např. pro tabulku víra podle pohlaví použijeme:

|         |   | víra |     | pohlaví | víra | frekvence |
|---------|---|------|-----|---------|------|-----------|
| pohlaví |   | 1    | 2   |         |      |           |
| 1       | 1 | 435  | 147 | 1       | 1    | 435       |
| 1       | 2 |      |     | 1       | 2    | 147       |
| 2       | 1 | 375  | 134 | 2       | 1    | 375       |
| 2       | 2 |      |     | 2       | 2    | 134       |

- stata syntax pro dvojrozměrnou tabulku

```
. tabulate pohlavi vira [w= freq]
(frequency weights assumed)
```

|         |   | vira |     | Total |
|---------|---|------|-----|-------|
| pohlavi |   | 1    | 2   |       |
| 1       | 1 | 435  | 147 | 582   |
| 2       | 1 | 375  | 134 | 509   |
| Total   |   | 810  | 281 | 1,091 |

## PRAVDĚPODOBNOST V KONTINGENČNÍ TABULCE

- základní typy pravděpodobnosti pro 2x2 tabulku jsou
  - **celková/sdružená pravděpodobnost** (pravděpodobnost výskytu jednotky v  $i$ -té variantě proměnné  $X$  a zároveň  $j$ -té variantě proměnné  $Y$ ), označení  $\pi_{ij}$  pro populaci a označení  $p_{ij}$  pro výběr (platí, že  $\sum \pi_{ij} = 1$ ,  $\sum p_{ij} = 1$ , výpočet  $p_{ij} = n_{ij} / N$ )
  - **marginální pravděpodobnost** (pravděpodobnost, že jednotka nabude  $i$ -té varianty  $X$  (nebo  $Y$ ) bez ohledu na  $Y$  (nebo  $X$ ), v tabulce jsou tyto pravděpodobnosti v posledním řádku nebo sloupci, označení  $p_{i+}$  ( $\pi_{i+}$ ) řádková proměnná,  $p_{+j}$  ( $\pi_{+j}$ ) sloupcová proměnná (platí  $p_{+1} = p_{11} + p_{12}$ , výpočet  $p_{+j} = n_{+j} / N$ )
  - **podmíněná pravděpodobnost** (relativní řádková, sloupcová pravděpodobnost), konstruuje se v případě, že rozlišujeme nezávisle (vysvětlující) a závisle (vysvětlovanou) proměnnou, např.  $Y$  podle  $X$ , jedná se o pravděpodobnost  $Y$  v každé variantě  $X$ , označení  $p_{ij}$  nebo  $p_{j/i}$  ( $\pi_{i/j}$ ,  $\pi_{j/i}$ ) (platí, že  $\sum p_{ij} = 1$ , výpočet např.  $p_{1/1} = n_{1/1} / n_{1/+}$ )
    - když je nezávisle proměnná v řádcích, počítáme podmíněnou pravděpodobnost v řádcích podle sloupců (**interpretace!**)
    - když je nezávisle proměnná ve sloupcích, počítáme podmíněnou pravděpodobnost ve sloupcích podle řádků (**interpretace!**)

## NEZÁVISLOST PROMĚNNÝCH V KONTINGENČNÍ TABULCE

- dvě proměnné X a Y jsou statisticky nezávislé tehdy, když podmíněná pravděpodobnost X (Y) je stejná v každé variantě Y (X)
- relativní řádková (sloupcová) pravděpodobnost je tedy v každém poli tabulky stejná
- např. víra v posmrtný život je nezávislá na pohlaví

| pohlaví | víra |         |
|---------|------|---------|
|         | ano  | ne/neví |
| žena    | 0.7  | 0.3     |
| muž     | 0.7  | 0.3     |

| pohlaví | víra |         |
|---------|------|---------|
|         | ano  | ne/neví |
| žena    | 0.5  | 0.5     |
| muž     | 0.5  | 0.5     |

- výpočet očekávaných četností v dvojrozměrné kontingenční tabulce:

$$F_{ij} = \frac{f_{i+} \cdot f_{+j}}{f_{++}}$$

- očekávané četnosti ukazují rozložení případů v tabulce za situace statistické nezávislosti mezi proměnnými X a Y

## NEZÁVISLOST PROMĚNNÝCH V KONTINGENČNÍ TABULCE

- pro test statistické nezávislosti mezi proměnnými v kontingenční tabulce se používá **Pearsonův chí-kvadrát test** ( $\chi^2$ ) se stupni volnosti  $(i - 1)(j - 1)$

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(F_{ij} - f_{ij})^2}{F_{ij}}$$

- dále se používá **Poměr maximální věrohodnosti** ( $L^2$ , někdy  $G^2$ ), či věrohodnostní poměr, se stejným počtem stupňů volnosti  $(i - 1)(j - 1)$

$$L^2 = 2 \sum_{i=1}^I \sum_{j=1}^J f_{ij} \ln \left( \frac{f_{ij}}{F_{ij}} \right)$$

- protože  $F_{ij} - f_{ij}$  nazýváme reziduály, měří  $\chi^2$  a  $L^2$  sednutí modelu nezávislosti na data, tedy odchylku očekávaných četností od pozorovaných, odpovídají na otázku, jak moc se model liší od dat? Obě tyto statistiky mají stejnou  $\chi^2$  distribuci, každá z nich má ovšem své výhody a nevýhody ( $\chi^2$  se používá spíše při souborech s malým  $N$ )



# NEZÁVISLOST PROMĚNNÝCH V KONTINGENČNÍ TABULCE

- **příklad:** pozorované četnosti a výsledky testu  $\chi^2$

| pohlaví | víra  |         |
|---------|-------|---------|
|         | ano   | ne/neví |
| žena    | 435   | 147     |
|         | 432.1 | 149.9   |
|         | 0.019 | 0.056   |
| muž     | 375   | 134     |
|         | 377.9 | 131.1   |
|         | 0.022 | 0.064   |

Pearson chi2(1) = 0.1621 Pr = 0.687

Likelihood-ratio chi2(1) = 0.1620 Pr = 0.687

Odhadnutý model nezávislosti se statisticky významně neliší od dat (df=1), proto tento model můžeme přijmout a konstatovat, že proměnné pohlaví a víra spolu nesouvisejí

- **adjustované reziduály (AR):** ukazují rozdíly mezi  $f_{ij}$  a  $F_{ij}$ , je to jedno číslo pro každé tabulkové pole, tyto čísla jsou mezi sebou komparovatelná (logika výpočtu: Pearsonův residuál  $(f_{ij} - F_{ij} / F_{ij}^2)$  dělený odhadnutou standardní chybou), cílem AR je lépe porozumět struktuře dat

$$AR = \frac{f_{ij} - F_{ij}}{\sqrt{F_{ij}(1 - p_{i+})(1 - p_{+j})}}$$

- hodnota  $AR > 2$  nebo  $3$  indikuje odmítnutí  $H_0$  o tom, že neexistují statistické rozdíly mezi  $f_{ij}$  a  $F_{ij}$  (jsou v mezích náhody)

## ASOCIACE V KONTINGENČNÍ TABULCE - RELATIVNÍ RIZIKO (RR)

- pro dichotomickou závisle proměnnou v kontingenční tabulce stačí znát pouze podíl  $p$  pro jednu variantu, pro druhou variantu je podíl  $(1-p)$ , je to doplnění do čísla 1
- obecně platí, že pro závisle proměnnou s  $J$ -kategoriemi je  $J-1$  podílů neredundantních
- relativní riziko (RR) je **poměr podmíněných podílů** (poměr relativních řádkových nebo sloupcových četností)

$$RR = \frac{p_{2|1}}{p_{1|1}}; \text{ zbytek } RR = \frac{p_{2|2}}{p_{1|2}} = \frac{(1 - p_{2|1})}{(1 - p_{1|1})}$$

- RR je vždy kladné číslo, 1 znamená nezávislost  $Y$  na  $X$  ( $p_{2|2} = p_{1|1}$ )
- **příklad:**

| pohlaví | souhlas s předmanž. sexem |              |
|---------|---------------------------|--------------|
|         | ano                       | ne           |
| žena    | 165<br>30.84              | 370<br>69.16 |
| muž     | 435<br>74.74              | 147<br>25.26 |

- RR muži =  $74.74 / 30.84 = 2.42$ ; RR pro muže oproti ženám pro variantu ano je **2.42 krát** větší, neboli o **142%** větší  $((2.42-1) \cdot 100)$
- RR ženy =  $30.84 / 74.74 = 0.41$ ; RR pro ženy oproti mužů pro variantu ne je **0.41 krát** menší, neboli o **59%** menší  $((1-0.41) \cdot 100)$
- číslo 1 v obou případech označuje nezávislost a čísla ukazují jednu a tu samou věc, ale naopak
- převod na přirozený logaritmus to dokazuje, protože  $\ln(2.43) = -\ln(0.41)$
- $\ln(2.43) = \mathbf{0.89}$ ;  $\ln(0.41) = \mathbf{-0.89}$
- číslo 0 v tomto případě označuje nezávislost, obě čísla jsou od 0 ve stejné vzdálenosti, ovšem v opačné směru
- např.  $(\underline{5} \times \mathbf{2.42}) = 12.1$  a  $(12.1 \times \mathbf{0.41}) = \underline{5}$

## ODDS RATIO (OR) - POMĚR ŠANCÍ

- OR ukazuje asociaci v kontingenčních tabulkách, OR je základním stavebním kamenem loglineárních modelů, OR jsou rovněž důležité pro pochopení logiky logistické regrese
  - RR je poměr dvou podmíněných pravděpodobností
  - OR je poměr dvou šancí (odds)
- **šance (O)** je poměr je poměr pravděpodobnosti jedné varianty (události) ke druhé variantě (událost nenastala)
- **příklad výpočtu šancí:**

| pohlaví | souhlas s předmanž. sexem |          |
|---------|---------------------------|----------|
|         | ano                       | ne       |
| žena    | $p_{11}$                  | $p_{12}$ |
| muž     | $p_{21}$                  | $p_{22}$ |

$$O_{(ne/ano)} = \frac{p_{12}}{p_{11}} = \frac{p_{12}}{(1 - p_{12})}$$

$$O_{(zeny/muzi)} = \frac{p_{21}}{p_{11}} = \frac{p_{21}}{(1 - p_{21})} \Rightarrow p = \frac{\text{Odd}}{\text{Odd} + 1}$$

- šance ukazuje pravděpodobnost, že se určitá událost stala, je to vždy kladné číslo
  - 1 znamená stejný výskyt, stejnou šanci pro obě konkurenční události
  - >1 vyšší šance pro událost (variantu)
  - <1 nižší šance pro událost (variantu)

## OR - POMĚR ŠANCÍ

- příklad výpočtu šancí:

- $O_{(\text{muži/ano})} = 2.64$  (2.64 krát větší šance pro ano u mužů oproti ženám, nebo 264 souhlasů u mužů ku 100 souhlasům u žen, nebo o 164% více pro ano u mužů)
- $O_{(\text{ženy/ano})} = 0.38$  (0.38 krát menší šance pro ano u žen oproti mužům, nebo 38 ano u žen na 100 ano u mužů nebo o 62% méně pro ano u žen)
  - 2.64 odpovídá 0.38 (důkaz - převod na přirozený logaritmus, 0 pak označuje stav nezávislosti)
  - tvrzení 2.64 krát více odpovídá tvrzení o 164% více (důkaz: zvolme libovolné přirozené číslo, např. 3, pak platí, že
    - (a)  $3 \times 2.64 = \underline{7.92}$  (dostáváme číslo, které je 2.64x větší než zvolené číslo 3)
    - (b)  $1\% \text{ z } 3 = 0.03$
    - (c)  $0.03 \times 164 = 4,92$
    - (d)  $3 + 4,92 = \underline{7,92}$  (dostáváme číslo, které je o 164% větší než zvolené číslo 3)
    - (e) výsledek rovnice (1) = výsledku rovnice (4)

## OR - POMĚR ŠANCÍ

- **OR** se vypočítá jako poměr dvou šancí (rozlišujeme pozorované OR nebo na základě očekávaných četností vypočítané (modelový) OR)

$$\theta = \frac{\omega_1}{\omega_2} = \frac{\frac{p_{11}}{p_{21}}}{\frac{p_{12}}{p_{22}}} = \frac{p_{11} \cdot p_{22}}{p_{21} \cdot p_{12}} = \frac{f_{11} \cdot f_{22}}{f_{21} \cdot f_{12}} \qquad \theta = \frac{\omega_2}{\omega_1} = \frac{\frac{\pi_{22}}{\pi_{21}}}{\frac{\pi_{12}}{\pi_{11}}} = \frac{\pi_{22} \cdot \pi_{11}}{\pi_{21} \cdot \pi_{12}} = \frac{F_{11} \cdot F_{22}}{F_{21} \cdot F_{12}}$$

- OR je kladné číslo, variuje v intervalu  $\langle 0; \infty \rangle$ , interpretace závisí na zvolené referenční kategorii,  $OR > 1$  nebo  $OR < 1$  znamená asociaci mezi variantami proměnných, čím větší vzdálenost od 1 tím také větší asociace,  $OR = 1$  znamená nezávislost
- 2 hodnoty OR u stejných kategorií reprezentují jednu a tu samou variantu asociace, ovšem v opačném směru (např.  $OR=4$  a  $OR=0.25$ )
  - kontrastní hodnotu asociace dostaneme  $1/OR$  ( $1/4=0.25$  nebo  $1/0.25=4$ ), interpretace je stejná jako u šancí (O) nebo u RR
  - LOR (log-odds-ratio) je přirozený logaritmus poměru šancí, variuje  $\langle -\infty; \infty \rangle$ , např.  $OR = 4$ , pak  $LOR = 1,39$  (nebo  $OR=0.25$ , pak  $LOR= -1.39$ )
  - **převod tabulkových četností na ln a výpočet OR!**
- **interpretace OR!**, je to vztah 2 šancí, ne poměrů nebo čísel

## OR - POMĚR ŠANCÍ

- OR se také někdy nazývá tabulkový poměr (cross-product ratio)
- pro 2x2 kontingenční tabulku existuje pouze 1 smysluplný poměr šancí, protože volba jiné referenční kategorie vede ke stejnému OR nebo jemu jinému číselnému vyjádření, které ovšem substantivně znamená stejnou věc
- obecně platí, že pro  $I \times J$  dimenze v tabulce stačí vypočítat  $(I-1)(J-1)$  poměru šancí, zbylé OR odvodíme z již vypočítaných OR

- obecně platí:

$$\theta_{ij} = \frac{F_{ij} \cdot F_{(i+1)(j+1)}}{F_{i(j+1)} \cdot F_{(i+1)j}} \quad (i = 1, \dots, I - 1; j = 1, \dots, J - 1)$$

- v  $I \times J$  tabulce je mnoho OR, protože každé OR zahrnuje kombinaci 2 řádkových variant jedné proměnné a 2 sloupcových variant druhé proměnné
- protože u OR jsou pojaty proměnné symetricky, není nezbytné při jejich výpočtu rozlišovat závisle a nezávisle proměnnou, u RR a jeho interpretaci to bylo nezbytné, protože hodnota RR závisela na tom, zdali jsem RR počítali v první nebo druhé variantě závisle proměnné
- vztah mezi OR a RR je:

$$OR = \frac{p_1(1-p_1)}{p_2(1-p_2)} = RR \left( \frac{1-p_1}{1-p_2} \right)$$

## OR - POMĚR ŠANCÍ

- OR jsou invariantní
  - k celkovému počtu případů (když změním velikost  $N$  o konstantu  $C$ , OR zůstává konstantní)

$$\theta = \frac{c \cdot f_{11} \cdot c \cdot f_{22}}{c \cdot f_{12} \cdot c \cdot f_{21}} = \frac{\cancel{c} \cdot f_{11} \cdot \cancel{c} \cdot f_{22}}{\cancel{c} \cdot f_{12} \cdot \cancel{c} \cdot f_{21}} = \frac{f_{11} \cdot f_{22}}{f_{12} \cdot f_{21}}$$

- k řádkové marginální distribuci (když změním první řádek o konstantu  $C$  a druhý řádek o konstantu  $D$ , OR zůstává konstantní)

$$\theta = \frac{c \cdot f_{11} \cdot d \cdot f_{22}}{c \cdot f_{12} \cdot d \cdot f_{21}} = \frac{\cancel{c} \cdot f_{11} \cdot \cancel{d} \cdot f_{22}}{\cancel{c} \cdot f_{12} \cdot \cancel{d} \cdot f_{21}} = \frac{f_{11} \cdot f_{22}}{f_{12} \cdot f_{21}}$$

- k sloupcové marginální distribuci (když změním první sloupec o konstantu  $C$  a druhý sloupec o konstantu  $D$ , OR zůstává konstantní)

$$\theta = \frac{c \cdot f_{11} \cdot d \cdot f_{22}}{d \cdot f_{12} \cdot c \cdot f_{21}} = \frac{\cancel{c} \cdot f_{11} \cdot \cancel{d} \cdot f_{22}}{\cancel{d} \cdot f_{12} \cdot \cancel{c} \cdot f_{21}} = \frac{f_{11} \cdot f_{22}}{f_{12} \cdot f_{21}}$$

- z tohoto důvodu se OR využívají především v těch případech, kdy je nutné odhlédnout od marginálních distribucí (např. při analýze mobilitních tabulek)

# Lekce 3:

## Analýza vícerozměrných tabulek v sociologii



## PARCIÁLNÍ A MARGINÁLNÍ KONTINGENČNÍ TABULKY

- vícerozměrné tabulky, problém asociace mezi proměnnými, otázka vztahu mezi dvěma proměnnými při kontrole třetí proměnné
- **parciální tabulky**
  - modelování dvojrozměrných tabulek podle třetí proměnné, zobrazujeme vztah mezi  $X$  a  $Y$  v jednotlivých variantách  $Z$ ,  $Z$  je drženo na stejné hladině, což znamená mapování vlivu  $X$  na  $Y$  při kontrole  $Z$ , je to **podmíněná asociace** mezi  $X$  a  $Y$ , protože je kontrolována pro  $Z$ 
    - otázka: zmizí vztah mezi  $X$  a  $Y$  při kontrole pro  $Z$ , nebo stále existuje?
- **marginální tabulky**
  - dvojrozměrné tabulky, nebereme zřetel na třetí nebo další proměnné, parciální tabulky jsou zkombinovány do dvojrozměrné marginální tabulky, každé pole v tabulce je pak sumou toho samého pole v jednotlivých parciálních tabulkách

## PARCIÁLNÍ A MARGINÁLNÍ KONTINGENČNÍ TABULKY

- Jaký je vztah mezi barvou pleti opakovaně obžalovaných z vraždy, uvalením trestu smrti na ně a barvou pleti jejich obětí? Neboli, jak poznamenává barva pleti u opakovaně obžalovaných rozhodnutí o jejich trestu smrti při kontrole barvy pleti obětí? (data pocházejí z amerického státu Florida, byla sebraná mezi lety 1976-1987, Agresti 1996, str. 56)

| Oběť    | Obžalovaný | Trest smrti |     | (% ) Ano |
|---------|------------|-------------|-----|----------|
|         |            | Ano         | Ne  |          |
| Běloši  | Běloši     | 53          | 414 | 11,3     |
|         | Černoši    | 11          | 37  | 22,9     |
| Černoši | Běloši     | 0           | 16  | 0        |
|         | Černoši    | 4           | 139 | 2,8      |
| Celkem  | Běloši     | 53          | 430 | 11       |
|         | Černoši    | 15          | 176 | 7,9      |

- co ukazuje marginální tabulka?
- co ukazují parciální tabulky?
- údaje z marginální tabulky vs. údaje z parciálních tabulek?
- OR z marginálního vztahu vs. OR z parciálního vztahu
- marginální asociace ukazuje opak parciální (podmíněné) asociace
  - tato skutečnost se nazývá **Simpsonův paradox**

## PODMÍNĚNÉ A MARGINÁLNÍ POMĚRY ŠANCÍ (OR)

- OR pro vícerozměrné tabulky počítáme jako podmíněné OR v jednotlivých variantách třetí proměnné

$$\theta_{xy|k} = \frac{f_{11|k} \cdot f_{22|k}}{f_{21|k} \cdot f_{12|k}} \quad \ln \theta_{xy|k} = \ln f_{11|k} + \ln f_{22|k} - (\ln f_{21|k} + \ln f_{12|k})$$

- OR pro marginální tabulku

$$\theta_{xy} = \frac{\frac{p_{11}}{p_{12}}}{\frac{p_{21}}{p_{22}}} = \frac{p_{11} \cdot p_{22}}{p_{21} \cdot p_{12}} = \frac{f_{11} \cdot f_{22}}{f_{21} \cdot f_{12}} \quad \ln \theta_{xy} = \ln f_{11} + \ln f_{22} - (\ln f_{21} + \ln f_{12})$$

## PODMÍNĚNÉ A MARGINÁLNÍ POMĚRY ŠANCÍ

- Dvě marginální tabulky, které ukazují jednak dvě nemocnice, které aplikovaly naprosto stejnou léčbu pro drogově závislé, a jednak výsledky, kterých dosáhly.

| Drogová léčba | Výsledek |    |
|---------------|----------|----|
|               | +        | -  |
| A             | 20       | 20 |
| B             | 20       | 40 |

| Nemocnice | Výsledek |    |
|-----------|----------|----|
|           | +        | -  |
| Praha     | 30       | 20 |
| Brno      | 10       | 40 |

| Nemocnice | Drogová léčba |    |
|-----------|---------------|----|
|           | A             | B  |
| Praha     | 30            | 20 |
| Brno      | 10            | 40 |

| Nemocnice | Drogová léčba | Výsledek |    |
|-----------|---------------|----------|----|
|           |               | +        | -  |
| Praha     | A             | 18       | 12 |
|           | B             | 12       | 8  |
| Brno      | A             | 2        | 8  |
|           | B             | 8        | 32 |
| Celkem    | A             | 20       | 20 |
|           | B             | 20       | 40 |

- proč je drogová léčba A dvakrát úspěšnější než drogová léčba B?
  - Praha častěji aplikuje léčbu A než Brno a zároveň má také pozitivní výsledky, parciální asociace tuto skutečnost ukazuje; závěr, že A je úspěšnější než B se při kontrole nemocnice ukáže jako falešný (rozdíl mezi A a B mizí)

## HOMOGENNÍ ASOCIACE

- homogenní asociace je konstantní asociace mezi dvěma proměnnými v jednotlivých variantách třetí proměnné (stejná velikost podmíněné asociace)
  - $OR_{xy(1)} = OR_{xy(2)} = \dots OR_{xy(k)}$
- když platí  $OR_{xy(1)} = OR_{xy(2)} = \dots OR_{xy(k)}$ , pak platí  $OR_{xz(1)} = OR_{xz(2)} = \dots OR_{xz(k)}$ ; a rovněž  $OR_{yz(1)} = OR_{yz(2)} = \dots OR_{yz(k)}$ 
  - homogenní asociace je vždy symetrická pro všechny varianty parciální asociace v jednotlivých variantách dalších proměnných Z, X nebo Y
- homogenní asociace znamená, že neexistuje trojrozměrná interakce, Z neovlivňuje vztah mezi X a Y
  - když homogenní asociace neexistuje, pak podmíněné OR variují podle třetí proměnné

## Lekce 4:

Lineární regresní model, zobecněné lineární modely (GLM), principy statistického modelování

## JEDNODUCHÁ LINEÁRNÍ REGRESE

- jednoduchá lineární regrese může být pro výběrový soubor zapsána jako:

$$\hat{y}_i = a + bx_i + d_i$$

- pro populaci je pak zapsána jako:

$$y_i = \alpha + \beta x_i + e_i$$

- kde  $\alpha$  je posunutí (intercept),  $\beta$  je směrnice pro jednotlivé varianty  $x$  a  $e$  je chyba (residuál, odchylka) pozorované proměnné od odhadnuté směrnice
- v regresní analýze je hodnota závisle proměnné specifikována jako součet lineárních efektů nezávisle proměnné (prediktora) a chyb (residuálů, odchylek, diferencí)

## JEDNODUCHÁ LINEÁRNÍ REGRESE

- stata syntaxt pro regresní model

```
regress price mpg headroom trunk weight length
```

| Source      | SS        | df        | MS         |                 |                      |          |
|-------------|-----------|-----------|------------|-----------------|----------------------|----------|
| -----+----- |           |           |            | Number of obs = | 74                   |          |
| Model       | 242096575 | 5         | 48419315.1 | F( 5, 68) =     | 8.38                 |          |
| Residual    | 392968821 | 68        | 5778953.25 | Prob > F =      | 0.0000               |          |
| -----+----- |           |           |            | R-squared =     | 0.3812               |          |
| Total       | 635065396 | 73        | 8699525.97 | Adj R-squared = | 0.3357               |          |
| -----       |           |           |            | Root MSE =      | 2403.9               |          |
| -----       |           |           |            |                 |                      |          |
| price       | Coef.     | Std. Err. | t          | P> t            | [95% Conf. Interval] |          |
| -----+----- |           |           |            |                 |                      |          |
| mpg         | -85.75773 | 83.60769  | -1.03      | 0.309           | -252.5943            | 81.07882 |
| headroom    | -710.1846 | 444.8546  | -1.60      | 0.115           | -1597.878            | 177.5089 |
| trunk       | 111.1498  | 109.9446  | 1.01       | 0.316           | -108.2411            | 330.5408 |
| weight      | 4.420511  | 1.165629  | 3.79       | 0.000           | 2.094535             | 6.746488 |
| length      | -108.0777 | 42.56471  | -2.54      | 0.013           | -193.0142            | -23.1411 |
| _cons       | 15552.1   | 6027.182  | 2.58       | 0.012           | 3525.049             | 27579.16 |
| -----       |           |           |            |                 |                      |          |



## ODHAD JEDNODUCHÉ LINEÁRNÍ REGRESE - OLS

- když známe vzorec pro regresi:

$$\hat{Y} = a + bX$$

- tak na základě metody nejmenších čtverců směrnici a posunutí vypočítáme podle vzorců:

$$b = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

- směrnice prochází vždy průměry obou proměnných, a proto posunutí (a) vypočítáme dosazením průměrů do upravené regresní rovnice

## PŘEDPOKLADY LINEÁRNĚ REGRESNÍHO MODELU

- odpovídající funkční podoba (linearita a normální rozložení)
- minimální výskyt odlehlých pozorování
- normální rozložení náhodných chyb (residuálů), **problém podoby podmíněné distribuce** (např. podmíněné zešikmení)
- homoskedasticita (konstantní variabilita) náhodných chyb (residuálů, složek), **problém tvaru podmíněné distribuce** (např. podmíněná špičatost)
- neexistence korelace mezi náhodnými chybami (residuály) a vysvětlujícími proměnnými
- neexistence multikolinearity

## STATISTICKÁ INFERENCE V LINEÁRNÍ REGRESI

- konfidenční interval  $CI(\beta)$ .....  $b \pm t^*SE_b$
- test hypotézy  $H_0: \beta = 0$ , výpočet  $t$  statistiky:

$$t = \frac{b}{SE_b}$$

- tabulkové kritické hodnoty  $t$  rozdělení

## ODHAD JEDNODUCHÉ LINEÁRNÍ REGRESE - MLE

- cílem MLE (maximálně věrohodného odhadu) je najít takovou hodnotu koeficientu (parametru), který nejméně generuje výběrová data
- výběrové hodnoty  $y_i$  jsou výsledkem pravděpodobnostní (hustotní) funkce  $f(y_i|\theta)$ , kde  $\theta$  je neznámý parametr, který generuje hodnoty  $y$  v populaci
- věrohodnostní funkce je pak součin pravděpodobností (hustot) jednotlivých  $y_i$ :

$$L = \prod_{i=1}^n f(y_i; \theta) \qquad \ln L = \sum_{i=1}^n \ln f(y_i; \theta)$$

- hledáme takový koeficient (obvykle sadu koeficientů) které maximalizují  $L$ , MLE tedy porovnává všechny možné regresní koeficienty a odpovídá na otázku, s jakou věrohodností generují naměřená data, numericky je snazší počítat s přirozeným logaritmem  $L$  (hledáme maximum  $\ln L$ , což odpovídá maximu  $L$ )
- k maximalizaci věrohodnostní funkce je nutné znát matematický vzorec pro náhodný proces generující data v populaci
- v případě lineární regrese musíme tedy přijmout předpoklad o rozložení  $y$  ve variantách  $x$ , (neboli předpokládat distribuci residuálů na základě určitého algoritmu)
- pro spojité znaky v regresi platí, že residuály jsou nezávislé, mají konstantní variabilitu  $\sigma^2$  a normální rozložení s  $\mu=0$ .

## ODHAD JEDNODUCHÉ LINEÁRNÍ REGRESE - ML

- u spojité závisle proměnné předpokládáme, že je generována na základě normálního rozložení (Gaussova distribuce)

- pravděpodobnost (hustota) je:  $p(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right)$  **pro**  $-\infty < y < \infty$

- střední hodnota je:  $\mu = \alpha + \beta x_i$

- dosazením a pro

parametry  $\alpha$  a  $\beta$  dostaneme: 
$$p(y_i | \alpha, \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2\right]$$

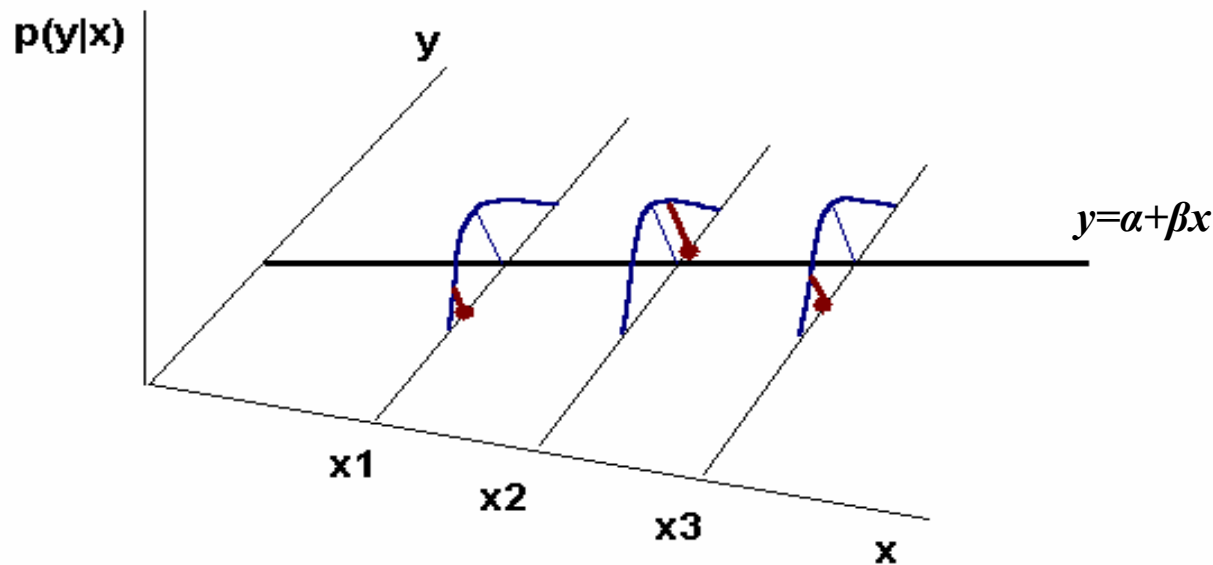
- $y_i$  a  $x_i$  jsou dány, zkoumáme pravděpodobnost pro varianty parametrů  $\alpha$  a  $\beta$
- věrohodnostní funkce  $L$  a přirozený logaritmus věrohodnostní funkce  $\ln L$ :

$$L(\alpha, \beta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2\right]$$

$$\ln L(\alpha, \beta) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2$$

## MLE A ZOBECNĚNÉ LINEÁRNÍ MODELY (GLM)

- $\ln L$  je součet všech pravděpodobností (hustot), která mají jednotlivé hodnoty  $x$  při daných parametrech
- cílem MLE je najít takové parametry, které maximalizují tento součet (je co nejbližší 0), v ideálním případě se tento součet = 0, všechna  $x$  leží na přímce a při daných parametrech mají  $\ln$  hodnoty pravděpodobnosti (hustoty) 0 ( $\exp^0 = 1$ )
- odhad koeficientů na základě MLE je totožný s odhade koeficientů na základě OLS, největší hodnoty  $\ln L$  dostaneme, když je v části vzorce  $(y_i - (\alpha + \beta x_i))^2$  rozdíl minimální ( $y_i$  se co nejvíce blíží  $\mu$ ), volíme tedy takové hodnoty parametrů  $\alpha$  a  $\beta$ , aby to platilo, OLS minimalizuje ten samý vztah, nicméně v termínech residuálů



## MLE A ZOBECNĚNÉ LINEÁRNÍ MODELY (GLM)

- když  $f$  je počet událostí z  $N$  pokusů (tedy pro pravděpodobnost  $y=1$ ) přijímáme předpoklad **binomického rozdělení**; po úpravě pro  $y=0$  (událost nenastala) a  $y=1$  (událost nastala) přijímáme **Bernoulliho rozdělení**
- když  $f$  je počet událostí v čase ( $y=1$ ), v místě nebo v rámci sociální skupiny (neznáme ovšem  $N$ , či počet událostí, které nenastaly ( $y=0$ )), přijímáme předpoklad **Poissonova rozdělení**
- všechna tato rozdělení patří do jedné rodiny distribucí (family), které matematicky vyjadřují náhodný proces, který generuje data (podle jejich typu)
- na základě těchto rozdělení a s pomocí spojnice (link) mezi závisle a nezávisle proměnnou lze tyto případy zobecnit
- hovoříme o zobecněných lineárních modelech (GLM)

## ZOBECNĚNÉ LINEÁRNÍ MODELY

- lineární prediktor  $v_i$  pro každou jednotku je:  $v_i = \mathbf{x}'_i \boldsymbol{\beta}$
- spojnice (*link function*)

$$\text{Identity: } \mu_i = v_i \qquad \text{Logit: } \mu_i = \frac{\exp(v_i)}{1 + \exp(v_i)} \Leftrightarrow \ln\left(\frac{\mu_i}{1 - \mu_i}\right) = v_i$$

$$\text{Log: } \mu_i = \exp(v_i) \Leftrightarrow \ln(\mu_i) = v_i \qquad \text{Probit: } \mu_i = \Phi(v_i) \Leftrightarrow \Phi^{-1}(\mu_i) = v_i$$

- podmíněné distribuce (*exponential family*):

$$\text{Gaussian: } y = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right)$$

$$\text{Binomial: } P(y) = \frac{N!}{y!(N-y)!} \pi^y (1 - \pi)^{N-y}$$

$$\text{Poisson: } P(y) = \frac{e^{-\mu} \mu^y}{y!}$$

**Gamma:**



## ZOBECNĚNÝ LINEÁRNÍ MODEL (GLM)

- jednoduchý regresní model je definován jako strukturní model:

$$Y_i = X_i' \beta + e_i$$

kde  $X_i$  je vektor hodnot pro  $i$ -té pozorování,  $\beta$  je vektor parametrů a  $e$  je chyba.

- statistický model je ve většině případů obsahuje:
  - **fixní část** (*fixed part, systematic component*), která popisuje vztah mezi proměnnými, které nás zajímají (tento vztah je obvykle lineární, a proto umožňuje zodpovědět otázku, jak proměnná  $X$  ovlivňuje  $Y$ )
  - **náhodná část** (*random part, random component*), jedná se o (reziduální) variaci vysvětlované proměnné, která je predikována na základě fixní části

## TYPY ZOBECNĚNÝCH LINEÁRNÍCH MODELŮ

| Fixní část     | Link            | Náhodná část | Model                           |
|----------------|-----------------|--------------|---------------------------------|
| spojitá        | identity        | normální     | regresní model                  |
| kategorizovaná | identity        | normální     | ANOVA                           |
| mix            | identity        | normální     | ANCOVA                          |
| mix            | logit           | binomická    | logistická regrese              |
| mix            | log             | poisson      | loglineární analýza             |
| mix            | zobecněný logit | multinomická | multinomická logistická regrese |

## ZOBECNĚNÉ LINEÁRNÍ MODELY (POKR.)

- stata syntaxt pro GLM

```
glm depvar varlist, family( ) link( )
```

kde

| <u>Family</u> | <u>Default</u> | <u>Link(spojnice)</u>   | <u>Other link</u> |
|---------------|----------------|-------------------------|-------------------|
| gaussian      | identity       | xb                      |                   |
| binomial      | logit          | $\exp(xb)/(1+\exp(xb))$ | probit, c-log-log |
| poisson       | log            | $\exp(xb)$              |                   |
| gamma         | log            | $\exp(xb)$              | 1/xb              |

## CO JE DOBRÝ STATISTICKÝ MODEL?

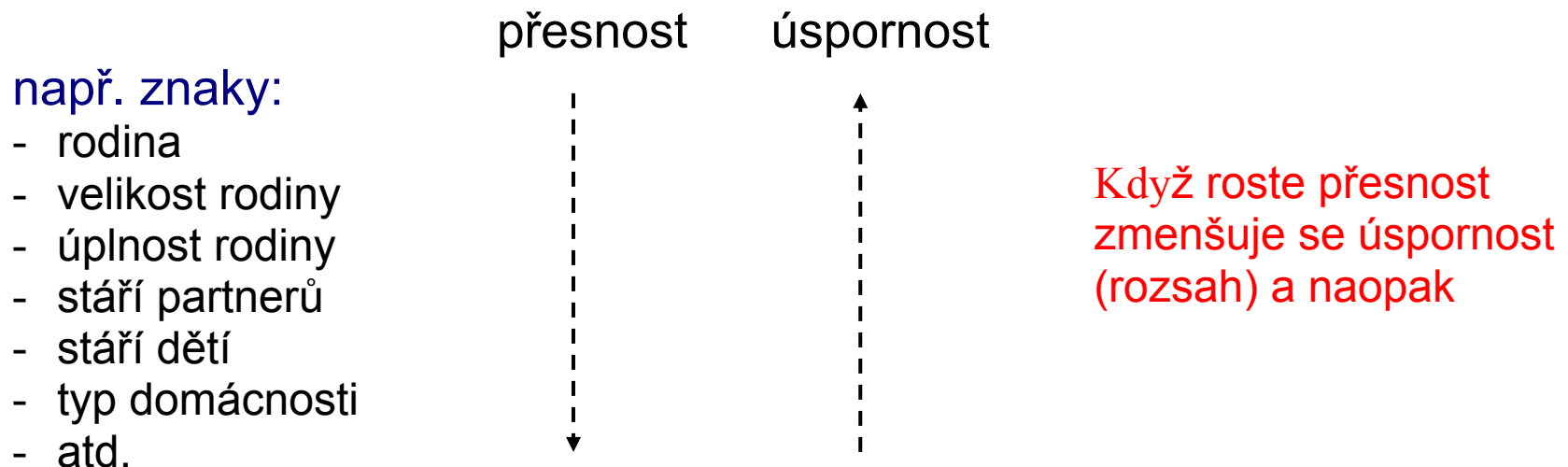
- dobrý statistický model je:
  - je **přesný** (očekávané četnosti, variabilita, podmíněný průměr) se co nejméně liší od pozorovaných četností, variability, podmíněného průměru)
  - je **úsporný** (obsahuje co nejméně parametrů, vysvětlujících proměnných)
    - koncept přesnosti (*accuracy*) = statistická kritéria  $X^2$ ,  $L^2$
    - koncept úspornosti (*parsimony*) = stupně volnosti (d.f. degrees of freedom)
- saturevaný model (obsahuje všechny vysvětlující proměnné a vztahy mezi nimi) **je přesný** (očekávané = pozorované četnosti,  $X^2$  a  $L^2 = 0$ ,  $df = 0$ ), ale **není úsporný** (je to parametrizace pozorovaných četnosti, nic nevysvětluje)
- model (podmíněné) nezávislosti (obsahuje obvykle minimum proměnných a vztahů mezi nimi), **je úsporný**, ale obvykle **není přesný** (rozdíl mezi očekávanými pozorovanými četnostmi je velký,  $X^2$  &  $L^2$  vysoké,  $df$  vysoké, málo parametrů na explanaci)

## PRINCIPY STATISTICKÉHO MODELOVÁNÍ

- v modelování výzkumník obvykle postupuje tak, že hledá model (v případě, že model (podmíněné) nezávislosti na data nepadne), který se nachází někde mezi saturevaným modelem a modelem nezávislosti
- modelování je hledání optimálního poměru mezi **přesností** a **úsporností** (logika Occamovy břitvy)
- cílem je najít co nejúspornější model, který má co nejméně vysvětlujících proměnných, který ovšem stále ještě uspokojivě vysvětluje strukturu dat
- důvod minimalizace vysvětlujících proměnných v modelu
  - numerická stabilita
  - snadná zobecnitelnost a aplikovatelnost
- dva možné postupy statistického modelování
  - začneme saturevaným modelem a postupně vylučujeme proměnné (snižuje se přesnost, ale roste úspornost) (*backward elimination in stepwise regression*)
  - začneme modelem (podmíněné) nezávislosti a postupně přidáváme proměnné (snižuje se úspornost, ale roste přesnost) (*forward addition in stepwise regression*),
  - v obou případech je kritériem pro proměnnou v modelu statistická významnost (obvykle 95%), problém hranice!
- **dobrá teorie je základem pro oprávněnost nebo neoprávněnost proměnných v modelu**

## VZTAH MEZI PŘESNOSTÍ A ÚSPORNOSTÍ V SCLG. VÝZKUMU

- každý zkoumaný (výběrový) soubor je definován **obsahem** a **rozsahem**
  - **obsah**: zkoumaný počet společných **znaků** u jednotek, konkrétnost, přesnost
  - **rozsah**: **počet** jednotek, úspornost
- větší obsah znamená větší počet znaků u jednotky, větší přesnost ve vymezení jednotky, nicméně to znamená vymezení menšího počtu jednotek (maximální počet znaků = 1 jednotka),
- větší rozsah, více zkoumaných jednotek, znamená menší počet znaků u jednotky (maximální rozsah = 1 znak) např. lidé



## REGRESNÍ MODELY PRO KATEG. ZÁVISLE PROMĚNNOU

- v případě kategorizované závisle proměnné regresní model nelze použít
- podle typu závisle proměnné volíme:
  - **binární logistickou regresi** - závisle proměnná má dvě varianty
  - **ordinální logistickou regresi** - závisle proměnná více uspořádaných variant
  - **nominální (multinomickou) logistickou regresi** - závisle proměnná více variant

### Shrnutí jednotlivých typů analýzy:

| Závisle proměnná         | Nezávisle proměnná      | Typ analýzy                |
|--------------------------|-------------------------|----------------------------|
| spojitá                  | spojitá                 | regrese, korelační analýza |
| spojitá                  | kategorizovaná          | regrese, ANOVA             |
| dichotomická (binární)   | kategorizovaná          | logit/probit, loglinear    |
| dichotomická (binární)   | spojitá                 | logit/probit               |
| neuspořádaná polytomická | kategorizovaná          | loglinear, mlogit          |
| neuspořádaná polytomická | spojitá                 | mlogit                     |
| uspořádaná polytomická   | kategorizovaná          | ologit/oprobit, loglinear  |
| uspořádaná polytomická   | spojitá                 | ologit/oprobit             |
| tabulková data (poměry)  | kategorizovaná          | loglinear                  |
| censored duration data   | spojitá, kategorizovaná | loglinear, logit/log-log   |