Komparační tabulky

Výchozí bod

- **Jednoduché četnostní tabulky kategoriálních proměnných**
- **Kombinace dvou a více kategoriálních znaků**
- **Úkoly:**
  - Porovnání pozorovaných četnosti u dvou a více znaků
  - Zjištění zda mezi znaky existuje závislost
  - Pokud existuje závislost, jaká je její síla

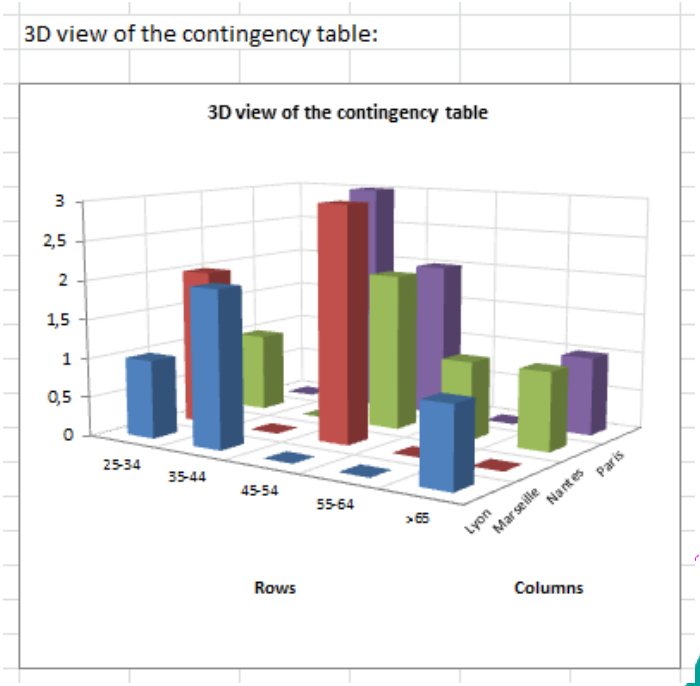

Schéma kontingenční tabulky - značení

Kontingenční tabulka je zápis o výskytu jevů v křížové kombinaci dvou kategorizací: řádkové A = (A_1, A_2, … A_R) a sloupcové B = B_1, B_2, … B_C)

A \ B	1	2	…	c	…	C	celkem
1	n_{11}	n_{12}	…	n_{1c}	…	n_{1C}	n_{1+}
2	n_{21}	n_{22}	…	n_{2c}	…	n_{2C}	n_{2+}
…			…		…	…	…
r	n_{r1}	n_{r2}	…	n_{rc}	…	n_{rC}	n_{r+}
…			…		…	…	…
R	n_{R1}	n_{R2}	…	n_{Rc}	…	n_{RC}	n_{R+}
celkem	n_{+1}	n_{+2}	…	n_{+c}	…	n_{+C}	n

- ***n* jsou absolutní četnosti výskytů**
- **zaměníme-li písmena *f* místo *n*, dostaneme analogický záznam o relativních četnostech, součet hodnot v tabulce 1 (místo *n*)**

Řádkové proporce

A \ B	1	2	…	c	…	C	celkem	celkové rozložení v řádcích
1	$f_{1/1}$	$f_{2/1}$	…	$f_{c/1}$	…	$f_{C/1}$	1	f_{1+}
2	$f_{1/2}$	$f_{2/2}$	…	$f_{c/2}$	…	$f_{C/2}$	1	f_{2+}
…	…	…	…	…	…	…	…	…
r	$f_{1/r}$	$f_{2/r}$	…	$f_{c/r}$		$f_{C/r}$	1	f_{r+}
…	…	…	…	…	…	…	…	…
R	$f_{1/R}$	$f_{2/R}$	…	$f_{c/R}$	…	$f_{C/R}$	1	f_{R+}
celkem	f_{+1}	f_{+2}	…	f_{+c}	…	f_{+C}	1	1

relativní četnosti $f_{c/r}$ dávají v součtu 1 v každém řádku (též v marginálním sloupci)

$f_{c/r}$, f_{+C}, f_{r+} jsou obvykle zaměňovány za 100* $f_{c/r}$%; místo 1 pak je v součtech 100%

Testování hypotézy o nezávislosti

- **nulová hypotéza - nezávislost řádkové a sloupcové proměnné**
  - **věta o součinu pravděpodobností: *P(současně A i B) = P(A)*P(B)***
  - **v buňce(C,R) tedy očekáváme, že podíl bude $f_{CR} = f_{+C} * f_{+R}$**
  - **očekávaný počet v buňce(c,r) je $e_{CR} = n_{1+} * n_{+2} / N$**
  - **rezidua – rozdíl mezi skutečným (*observed*) a očekávaným (*expcected*) počtem**
  - **testové kritérium založeno na reziduích**
- **použitelné testy**
  - **Pearsonův χ^2 test o nezávislosti**
  - **likelihood ratio**
  - **rozdělení χ^2 s (R-1)*(C-1) stupňů volnosti**

$$\chi^2_P = \sum\sum \frac{(o_{CR} - e_{CR})^2}{e_{CR}}$$

$$\chi^2_{LR} = -2 \sum\sum o_{CR} \ln(e_{CR}/o_{CR})$$

- **podmínky použití**
  - **očekávané četnosti pod 5 maximálně v 1/5 buněk**

Měření intenzity vztahu

- **míry pro číselné vyjádření síly závislosti**
- **Cramérovo V**
  - **vychází ze statistiky χ^2**
  - **interval 0 – 1, bez závislosti až po silnou závislost**
- **Koecifient φ**
  - **vychází ze statistiky χ^2**
  - **maximální hodnota (q-1)^1/2 - horší interpretace**
  - **čím vyšší, tím silnější závislost**
- **Koecifient kontingence**
  - **vychází ze statistiky χ^2**
  - **maximální hodnota (1-1/q)^1/2 - horší interpretace**
  - **čím vyšší, tím silnější závislost**

$q = \min(R,C)$

$$V = \sqrt{\frac{\chi_P^2}{n(q-1)}}$$

$$\phi = \sqrt{\frac{\chi_P^2}{n}}$$

$$CC = \sqrt{\frac{\chi_P^2}{\chi_P^2 + n(q-1)}}$$

Další míry síly závislosti

- Označují se jako predikční míry
- Dělí se podle použití kategoriální proměnné
- Odstraňují nevýhody měr vycházejících z chí - kvadrátu
- Nejznámější nominální míry
  - Goodmanův koeficient lambda
  - Goodman - Kruskalovo tau

- Nejznámější ordinální míry
  - Somerův koeficient
  - Koeficient gamma