



SEMINÁŘ VÝPOČETNÍ STATISTIKY

P12
2008-05-05

KVALITATIVNÍ PROMĚNNÉ (KATEGORIÁLNÍ):

Příklad:

- ✓ **Zadání:** Bylo zkoumáno, zda použití určitého očkovacího séra může snížit počet onemocnění nakažlivou chorobou. Pokus byl proveden u 23 pokusných zvířat stejného stáří (12 jich bylo očkováno, 11 neočkováno), když byla vystavena stejné nákaze. Výsledky šetření jsou uvedeny v následující tabulce:

Počet	Nakažených	Nenakažených	Celkem
Očkovaných	1	11	12
Neočkovaných	7	4	11
Celkem	8	15	23

✓ **Očekávané četnosti:**

- Zda nějaká očekávaná četnost neklesne pod hodnotu 5.

$$o_{ij} = \frac{12 \times 8}{23} = 4,17 < 5 - \text{není použitelný } \chi^2\text{-test pro ověření nulové hypotézy, která říká, že výskyt nákazy není závislý na očkování.}$$

- Použijeme tedy Fisherův test

✓ **Procedura** – lze výstup do .rtf, lze potlačit tisk některých věcí (norow, nocol, nopercent):

```
ods rtf;
proc freq data=svs;
tables ockovani*nakaza/norow nocol nopercent chisq measures;
weight pocet;
run;
ods rtf close;
```

✓ **Výstup:**

- Výstup rtf je graficky zdařilejší.
- Zobrazuje se asociační tabulka.
- Statistiky – je uvedeno varování, že 50% buněk má očekávané četnosti menší než 5 a chí-kvadrát test může být neplatný.
- Automaticky je uveden Fisherův test, uvedena dvoustranná p-hodnota, která je 0,0094, tedy není pravda, že by výskyt nákazy nebyl ovlivňován očkováním.
- Další tabulka zobrazuje statistiky gama či lambda (v asymerické a symetrické verzi). Nutno posoudit, zda znaky jsou nominální či ordinální a zda nás v případě nominálních zajímá asymetrický či symetrický tvar. Zajímá nás pouze vazba (rozlišení směru) nebo závislost v určitém směru? Nejspíše nás zajímá asymetrická závislost, tedy jestli je nemoc ovlivněna očkováním. C je sloupec, R je řádek, zajímá nás tedy asymetrická lambda C/R. Hodnota je 0,375.

Příklad:

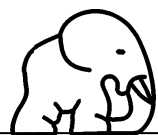
- ✓ **Zadání:** V určitém podniku byla prováděna analýza, zda příčina nespokojenosti pracovníků v daném podniku souvisí se stupněm jejich dosaženého vzdělání.

Tabulka pro vzdělání podle nespokojenosti						
Vzdělání	Nespokojenost					Součet
Četnost	Plat	Prostředí	Vztahy	Řízení	Jiné	
V	15	7	6	16	6	50
S	28	29	27	17	19	120
Z	38	32	28	12	20	130
Součet	81	68	61	45	45	300

- ✓ Nulová hypotéza říká, že dva kvalitativní znaky jsou na sobě nezávislé, tedy neexistuje žádná vazba.

✓ **Procedura:**

```
proc freq data=svs;
tables vzdelani*nespokojenost/norow nocol nopercent chisq measures;
weigh pocet;
run;
```

✓ **Výstup:**

- Tabulka nemá původní vzhled, došlo ke zpřeházení řádků a sloupců. Sas řadí sloupce a řádky abecedně.
- Lze však změnit přidáním příkazu do procedury a je možné mít výstup v rtf:

```
ods rtf;  
proc freq data=svs order=data;  
tables vzdelani*nespokojenost/norow nocol nopercent chisq measures;  
weigh počet;  
run;  
ods rtf close;
```

✓ **Výstup:**

- Chí-kvadrát můžeme použít, jelikož sas nezobrazil varování o četnostech.
- Pro posouzení síly závislosti Cramerův koeficient, hodnota 0,1733, závislost na vzdělání je tedy slabá.
- Vzdělání lze považovat za ordinální znak, ale další znak je nominální a využijeme tedy koeficient lambda. Volíme asymetrickou verzi, jelikož nás zajímá určitý směr, tedy asymetrické C/R, jelikož vzdělání je řádkové (row), hodnota je 0,0091.

ANALÝZA ČASOVÝCH ŘAD:

- ✓ Kopíruje problematiku korelační a regresní analýzy.

✓ **V SASu lze dvojím způsobem:**

- Série **procedur** s využitím jazyka SAS, dovoluje vytvořit analýzu na míru příslušnému problému, ale je to postup velmi náročný
- Komponenta systému, která automatizuje generování příslušných modelových postupů časových řad a z nabídky možných přístupů automaticky vybírá nejvhodnější přístup a bude generovat extrapolaci předpovědi časových řad. Snaha navrhnout výstižné předpovědi pro nějaké období dopředu.

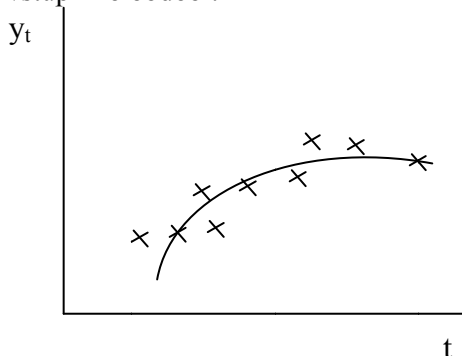
Komponenta se nazývá **TSFS** = Time Series Forecasting System.

Umožňuje modelování různými prostředky, nabízí přes 40 modelů a umožňuje podle vlastního uvážení modely doplňovat a nastavovat.

- ✓ Máme k dispozici **data**:

y_1, y_2, \dots, y_n
1 2 ... n

- ✓ **Časová řada** se zobrazí v grafu a připomíná korelační pole. Prokládáme vhodnou čarou, která co nejlépe vystihuje průběh řady v minulosti a pomocí funkce se snažíme provést prognózu, tedy prodloužit za okruh vstupního období.



- Časový úsek t je tzv. **referenční období**

✓ **Trendové modely:**

- **Klasické** – vhodné, když řada připomíná některou z funkcí:
 - Lineární
 - Kvadratický
 - Kubický
 - Exponenciální
 - Logaritmický
 - Apod.
- **Adaptivní** – modelují po částech, rozdělí na úseky a každý úsek se vyrovnává jiným způsobem. Je větší šance, že se pomocí nich podaří vyrovnat a extrapolovat časovou řadu.