SEMINÁŘ VÝPOČETNÍ STATISTIKY P3

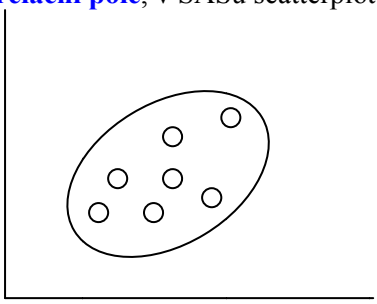
2008-03-03

VÍCEROZMĚRNÉ STATISTICKÉ SOUBORY:

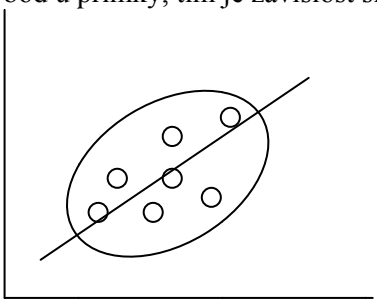
✓ Připadá v úvahu rozhodování, zda dvojice proměnných nejsou v nějakém vztahu.

Grafické znázornění:

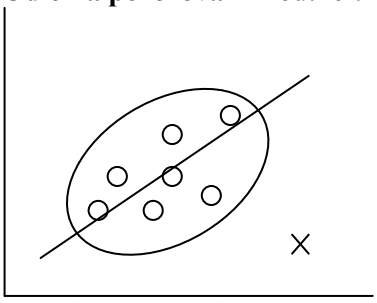
✓ **Korelační pole**, v SASu scatterplot:



- Přímka procházející korelačním polem – **regresní funkce**. Stanovuje, jak silná je závislost – čím blíže je bod u přímky, tím je závislost silnější



- **Odlehlá pozorování** – outlier:



✓ **Matice korelačních polí** – scatterplot matrix:

- Proměnné:

$(x_1, x_2, ...x_k)$

$(x_1, x_2)(x_1, x_3)...(x_1, x_k)$

$(x_2, x_1)(x_2, x_3)...(x_2, x_k)$

...

✓ **Příklad:**

Student	Hodiny	IQ	Body
1	9	99	56
2	5	100	45
3	12	119	80
4	14	95	73
5	11	110	71
6	6	117	55
7	19	98	95

8	16	101	86
9	3	100	34
10	9	115	66

- Tři proměnné, zjištění, zda mezi nimi existuje nějaká závislost (vazba).
- **Výstup** – pole na diagonále prázdná, jelikož by to byla závislost mezi jednou a toutéž veličinou. První je IQ, na ose y je IQ, na ose x u prvního grafu hodiny a u druhého body – porovnává se závislost IQ a hodin a IQ a bodů. Ostatní grafy stejné. Na diagonále v rozích dvě hodnoty – minimum a maximum.
- **Grafy** – jsou-li body hodně rozptýleny nesystematicky a nepřipomínají žádnou funkci, lze říci, že se mezi proměnnými neprojevuje závislost. Vytvářejí-li body systematickou strukturu, lze usoudit, že je mezi proměnnými lineární závislost. Jsou-li navíc body hodně zkoncentrovány, bude to ukazovat na silnou závislost.
- Zvolíme-li si jeden **bod**, zobrazí se zvýrazněně ve všech grafech. Objeví se číslo, které značí pořadí, tedy např. 7 znamená, že to bylo 7. pozorování. Pokud na bod klikneme opakovaně, objeví se u pozorování celá datová informace, tzn. vektor údajů.
- **Brushing** – zvolíme si graf a v něm klikneme do pole a táhneme polem, vytvoříme obdélníček (výběr). Všechny body se zvýrazní a zvýrazní se i v ostatních polích.
- **Matice korelačních polí** se vytváří v SAS Insight – spustit zapsáním tohoto slova do „vyhledávacího“ políčka a Enter. Vybereme jméno datového souboru a označíme si jména proměnných v souboru, která chceme prezentovat. Osy grafů nejsou pojmenovány, ale stačí kliknout v 1. poli na šipku a nabídky vybrat na „axes“.
- **Kauzální závislost** – příčinná, automaticky jsme nakloněni k tomu, abychom se domnívali, že závislost mezi veličinami skutečně existuje.

SASUser.Fitness:

✓ **Konstrukce histogramu a hustoty normálního rozdělení** (histogram) proměnné OXYGEN uložené v souboru SASUSER.FITNESS

- **Příkazy:**
  - Histogram – vytvoří histogram pro analyzovanou proměnnou
  - Normal – tento příkaz zabuduje do histogramu křivku hustoty normálního rozdělení
  - Color= – specifikace barvy křivky hustot, musí být uvedena v závorce
  - Cbarline= – specifikace barvy obrysů sloupků histogramu
  - Cfill= – určení barvy plochy jednotlivých sloupků histogramu
  - Noprint:
    - Potlačí tisk jednotlivých shrnujících charakteristik a testů, které vytváří procedura UNIVARIATE
    - Potlačí tisk shrnujících charakteristik zkonstruované hustoty normálního rozdělení (příkaz musí být uveden v závorce);
- **Procedura:**

```
proc univariate data=sasuser.fitness noprint;
var oxygen;
histogram oxygen/normal (noprint color=red) cbarline=blue cfill=yellow;
run;
```

✓ **Textový výstup:**

- **Procedura** – vynechat „noprint“ ze závorky:

```
proc univariate data=sasuser.fitness noprint;
var oxygen;
histogram oxygen/normal (color=red) cbarline=blue cfill=yellow;
run;
```

- Testy, které umožňují posoudit, do jaké míry se Gaussova křivka shoduje s histogramem, do jaké míry vystihuje křivka tvar histogramu. Klíčová je „p hodnota“, většinou vybíráme Kolmogorov-Smirnovův test. Test říká, že Gaussova křivka je vhodná. Pokud je p-hodnota menší než 5%, je potřeba testovanou hypotézu zamítnout. P-hodnota vyšla 0,119 => nemáme důvod hypotézu, která říká, že Gaussova křivka je vhodným vyrovnávacím nástrojem, zamítnout.

✓ **Neparametrická (jádrová) hustota:**

- Křivka, která se snaží maximálně přiblížit tvaru histogramu.
- Příkaz kernel – konstukce tzv. jádrové hustoty

- **Procedura:**
```
proc univariate data=sasuser.fitness noprint;
var oxygen;
histogram oxygen/cbarline=blue cfill=yellow kernel (color=green);
run;
```
- **Křivka jádrové hustoty** – méně pravidelná než Gaussova křivka, snaží se postihnout různé zlomy v histogramu. Křivka má hlavní vrchol, ale jsou naznačeny také dva menší (vedlejší) vrcholy, které ukazují na to, že datový soubor nemusel být homogenní, ale že data byla rozdělena do dvou skupinek, která se ze souboru vydělují.

✓ Chceme-li **obě křivky v jednom grafu** – procedura:
```
proc univariate data=sasuser.fitness noprint;
var oxygen;
histogram oxygen/normal (noprint color=red cbarline=blue cfill=yellow kernel
(color=green);
run;
```

✓ Průzkumová analýza rozdělení četnosti proměnné OXYGEN uložené v souboru SASUSER.FITNESS.

- **Příkazy:**
  - Mu0=50 – tímto příkazem je požadováno provedení testu hypotézy, že průměr základního souboru statistického znaku OXYGEN je roven 50
  - Cibasic – Výpočet intervalů spolehlivosti pro základní statistické charakteristiky
  - Type=lower – výpočet pouze dolní hranice těchto intervalů spolehlivosti, levostranné, musí být uvedeno v závorce. Pouze horní (pravostranné): Type=upper
  - Normal – výpočet testů normality rozdělení
  - Plot – konstrukce jednoduchých vizualizačních prostředků (stem-and-leaf display, boxplot, gplot)
- **Procedura:**
```
proc univariate data=sasuser.fitness mu0=50 cibasic (type=lower) normal plot;
var oxygen;
run;
```
- **Výstupy** – p-hodnoty nízké, hypotézu musíme zamítnout
  - Výběr testu:
    - Studentovo t – jednovýběrový t-test, pokud má výběr normální rozdělení.
    - Znamékový test – neparametrický test, znaménkový
    - Wilcoxonův test – neparametrický, v českém sasu Znam. pořadí, kvalitou se blíží jednovýběrovému t-testu